DOCUMENT RESUME

ED 356 275 TM 019 721

AUTHOR Kavanagh, Michael J.
TITLE Performance Rating Accuracy Improvement through Changes in Individual and System Characteristics.
INSTITUTION State Univ. of New York, Albany. Research Foundation.; Texas Maxima Corp., San Antonio.
SPONS AGENCY Air Force Human Resources Lab., Brooks AFB, TX. Training Systems Div.
REPORT NO AFHRL-TP-87-67
PUB DATE Apr 89
CONTRACT 85-004-12000-002; F33615-83-C-0030
NOTE 131p.
PUB TYPE Information Analyses (070) -- Reports - Research/Technical (143) -- Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC06 Plus Postage.
DESCRIPTORS *Evaluation Methods; Evaluators; *Graduate Students; Higher Education; *Individual Differences; *Job Performance; Motivation; Personnel Directors; *Personnel Evaluation; Standards; *Undergraduate Students
IDENTIFIERS Accuracy; *Performance Based Evaluation; System Evaluation

ABSTRACT

Although the quest for better measurement of individual job performance has generated considerable empirical research in industrial and organizational psychology, the feeling persists that a good job is not really being done in measuring job performance. This research project investigated the effects of differences in both individual and systems characteristics on the accuracy of job performance measurements using rating of individual effectiveness in fulfilling job duties. The research involved 4 studies over a period of 13 months. Subjects included 134 graduate students, 8 human resources managers, and 201 undergraduate students. Results indicate that: (1) the purpose for which performance ratings are collected does not affect accuracy; (2) the quality of the instructions that accompany the rating form can affect rating accuracy; and (3) the use of performance standards on the rating form and their effect on rating accuracy depend on the method used to collect the performance ratings. In addition, rater motivation, acceptance, and confidence are related to rating accuracy. The investigation of the methodologies used to collect the accuracy data suggests the need for new methods in future studies of rating accuracy. Eleven tables present study results, and 19 figures illustrate the discussion. Ten appendixes contain study questionnaires and instructions, and supplemental information about the studies. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

AFHRL-TP-87-67

AIR FORCE HUMAN RESOURCES LABORATORY

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it

☐ Minor changes have been made to improve reproduction quality

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

ED356275

PERFORMANCE RATING ACCURACY IMPROVEMENT THROUGH CHANGES IN INDIVIDUAL AND SYSTEM CHARACTERISTICS

Michael J. Kavanagh

School of Business
State University of New York at Albany
Albany, New York 12222

TRAINING SYSTEMS DIVISION
Brooks Air Force Base, Texas 78235-5601

April 1989

Final Technical Paper for Period October 1984 - December 1987

Approved for public release; distribution is unlimited.

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5601

TM019721

BEST COPY AVAILABLE

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.

HENDRICK W. RUCK, Technical Advisor
Training Systems Division

RODGER D. BALLENTINE, Lt Col, USAF
Chief, Training Systems Division

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S) AFHRL-TP-87-67
6a. NAME OF PERFORMING ORGANIZATION Texas Maxima Corporation 6b. OFFICE SYMBOL (*If applicable*)	7a. NAME OF MONITORING ORGANIZATION Training Systems Division
6c. ADDRESS (*City, State, and ZIP Code*) 8303 Broadway Suite 212 San Antonio, Texas 78209	7b. ADDRESS (*City, State, and ZIP Code*) Air Force Human Resources Laboratory Brooks Air Force Base, Texas 78235-5601
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Air Force Human Resources Laboratory 8b. OFFICE SYMBOL (*If applicable*) HQ AFHRL	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER F33615-83-C-0030
8c. ADDRESS (*City, State, and ZIP Code*) Brooks Air Force Base, Texas 78235-5601	10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
62703F	7734	08	24

11. TITLE (*Include Security Classification*)
Performance Rating Accuracy Improvement Through Changes in Individual and System Characteristics

12. PERSONAL AUTHOR(S)
Kavanagh, M.J.

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (*Year, Month, Day*)	15. PAGE COUNT
Final	FROM Oct 84 TO Dec 87	April 1989	110

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (*Continue on reverse if necessary and identify by block number*)
FIELD	GROUP	SUB-GROUP	job performance performance measurement rating accuracy
05	08		
05	09		

19. ABSTRACT (*Continue on reverse if necessary and identify by block number*)

The quest for better measurement of individual job performance has generated considerable empirical research in Industrial/Organizational Psychology; however, the feeling persists that we are not "really" doing a good job in measuring job performance. This research project was concerned with investigating the effects of differences in both individual and systems characteristics on the accuracy of job performance measurements using ratings of individual effectiveness in fulfilling job duties. The research involved four studies over a period of 13 months. Results indicated that: (a) the purpose for which the performance ratings are collected does not affect accuracy; (b) the quality of the instructions that accompany the rating form can affect rating accuracy; and (c) the use of performance standards on the rating form and their effect upon rating accuracy depend on the method used to collect performance ratings. In addition, rater motivation, acceptance, and confidence were found to be related to rating accuracy. Finally, the present investigation of the methodologies used to collect accuracy data suggested the need for new methods in future studies of rating accuracy.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION	
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	Unclassified	
22a. NAME OF RESPONSIBLE INDIVIDUAL Nancy J. Allin, Chief, STINFO Branch	22b. TELEPHONE (*Include Area Code*) (512) 536-3877	22c. OFFICE SYMBOL AFHRL/SCV

DD Form 1473, JUN 86 *Previous editions are obsolete.*

AFHRL Technical Paper 87-67 April 1989

PERFORMANCE RATING ACCURACY IMPROVEMENT THROUGH CHANGES IN INDIVIDUAL AND SYSTEM CHARACTERISTICS

Michael J. Kavanagh

School of Business
State University of New York at Albany
Albany, New York 12222

TRAINING SYSTEMS DIVISION
Brooks Air Force Base, Texas 78235-5601

Reviewed and submitted for publication by

Nestor K. Ovalle, II, Lt Col, USAF
Training Assessment Branch

This publication is primarily a working paper. It is published solely to document work performed.

SUMMARY

This research was conducted as part of the Air Force Job Performance Measurement (JPM) Project. The purpose was to evaluate the influence of four factors on the accuracy with which people rate the individual job performance of employees. In addition, the influence of four personal characteristics of raters on rating accuracy was addressed. Finally, two methodological issues arising from past research on rating accuracy were examined. The work was completed through four research studies conducted over a 13-month period.

Results indicate that rating accuracy: (a) is not affected by the purpose for collecting performance ratings; (b) is impacted by the quality of the instructions that accompany the rating form; and (c) is influenced by the use of performance standards on the rating form, although this depends on the method used to collect the performance ratings. Psychometrically sound measures for three of the four personal characteristics of raters (motivation, acceptance of the rating process, and confidence in ability to make accurate ratings) were developed and demonstrated that these characteristics were related to rating accuracy. The results of these studies indicate that procedures for future research on rating accuracy should be carefully established, since the present findings cast serious doubt on much of the previous research in the field of rating accuracy.

Finally, the results of this research provide specific guidelines and recommendations for other JPM project efforts.

PREFACE

This paper reports on four research studies done as part of an effort to develop a job performance measurement system (JPMS) for use by the Air Force in validating the Armed Services Vocational Aptitude Battery (ASVAB). Some practical issues regarding rating accuracy were evaluated, such as how to collect individual performance ratings, and the type of instructions that should accompany the rating form. Additionally, measures of important individual rater characteristics were developed and used to evaluate the impact personal attributes have on rating accuracy. Several methodological concerns were investigated as well. Specific recommendations for future JPMS research are given.

The work was performed by the Research Foundation of State University of New York, under subcontract 85-004-12000-002 with the MAXIMA Coporation and prime contract F33615-83-C-0030 (Task 12) from the Air Force Human Resources Laboratory (AFHRL) Manpower and Personnel Division. Dr. Michael J. Kavanagh was the Project Director. Barbara B. Kavanagh was the Project Administrator and Associate Scientist. She helped in project design, project administration, and data analysis. Thomas Lee was the Research Associate on this project. He assisted in data collection and data analyses. Dr. Jerry Hedge was the AFHRL Contract Monitor.

TABLE OF CONTENTS

Page

I. INTRODUCTION 1

General Background for Project 1
Research Variables 4
Literature Review 5

Purpose of Measurement 5
Acquaintance with the Job 6
Performance Standards 7
Quality of Instructions 7
Methodological Issues 8
Intervening Variables 9

Research Hypotheses 9

II. STUDY 1 16

Method 16

Experimental Design 16
Subjects 17
Stimulus Material 17
Research Variables 18
Procedure 19

Results 21

Manipulation Checks 21
Intervening Variables 21
ANOVA Analyses 21
Correlational Analyses 22

Discussion 24

III. STUDY 2 27

Method 27

Participants 27
Procedure 27

Results 28

Criterion Deficiency 28

Table of Contents (Continued)

Page

Performance Standards 29
SME-Derived True Scores 29

Discussion 29

IV. STUDY 3 32

Method 33

Experimental Design 33
Subjects 33
Research Variables 33
Experimental Procedure 34

Results 35

Intervening Variables 35
MANOVA Results 36
ANOVA Results 36
Correlational Results 40

Discussion 41

V. STUDY 4 43

Method 43

Experimental Design 43
Subjects 44
Research Variables 44
Experimental Procedure 45

Results 46

Intervening Variables 46
MANOVA Results 46
ANOVA Results 46
Correlational Results 47

REFERENCES 50

APPENDIX A: BIOGRAPHICAL QUESTIONNAIRE: STUDY 1 55

APPENDIX B: INSTRUCTIONS TO SUBJECTS: STUDY 1 57

Table of Contents (Concluded)

	Page
APPENDIX C: EXPERIMENTAL QUESTIONNAIRE: STUDY 1	71
APPENDIX D: INSTRUCTIONS TO SUBJECT MATTER EXPERTS	76
APPENDIX E: LOW LEVEL OF DETAIL INSTRUCTIONS	80
APPENDIX F: MODERATE LEVEL OF DETAIL INSTRUCTIONS	81
APPENDIX G: HIGH LEVEL OF DETAIL INSTRUCTIONS	83
APPENDIX H: BIOGRAPHICAL INFORMATION AND QUESTIONNAIRE	85
APPENDIX I: PERFORMANCE STANDARDS RATING FORM	90
APPENDIX J: POST-EXPERIMENTAL QUESTIONNAIRE	97

LIST OF FIGURES

Figure		Page
1	A Job Performance Measurement Classification Scheme	2
2	Descriptive Model for Rating Accuracy Project	3
3	Structural Model and Equations: Purpose of Measurement	10
3a	Model with Signed Relationships: Operational Purpose	10
3b	Model with Signed Relationships: Validation Purpose	11
4	Structural Model and Equations: Acquaintance with Job	12
4a	Model with Signed Relationships: Low Acquaintance with Job	12
4b	Model with Signed Relationships: High Acquaintance with Job	12
5	Structural Model and Equations: Peformance Standards	13
5a	Model with Signed Relationships: BARS Format	13
5b	Model with Signed Relationships: Performance Standards Format	14
6	Structural Model and Equations: Quality of Instructions	14

List of Figures (Concluded)

Figure		Page
6a	Model with Signed Relationships: Small Amount of Detail	15
6b	Model with Signed Relationships: Moderate Amount of Detail	15
6c	Model with Signed Relationships: Large Amount of Detail	15
7	Interaction for Distance Accuracy (SME): Study 3	38
8	Interaction for Correlational Accuracy (SME): Study 3	39
9	Interaction for Correlational Accuracy (Borman): Study 3	39
10	Interaction for Correlational Accuracy (Borman): Study 4	47

LIST OF TABLES

Table		Page
1	Reliabilities for Intervening Variables for Four Studies	21
2	Means for Significant Findings: Study 1	22
3	Correlation Results for Study 1	23
4	Intended Performance True Scores	30
5	Actual Performance True Scores	30
6	Subject-Matter Expert Performance True Scores	31
7	Means for Significant Findings for Intervening Variables: Study 3	36
8	Means for Significant Findings for Dependent Variables: Study 3	37
9	Correlation Results for Study 3	40
10	Means for Findings: Study 4	47
11	Correlation Results for Study 4	48

PERFORMANCE RATING ACCURACY IMPROVEMENT THROUGH CHANGES IN INDIVIDUAL AND SYSTEMS CHARACTERISTICS

I. INTRODUCTION

This research and development (R&D) effort investigated the effects of differences in both individual and system characteristics on the accuracy with which individuals rate the job performance of others. Specifically, it tested a subset of the hypothesized relationships in the performance measurement quality model (Figure 1) developed by Kavanagh, Borman, Hedge, and Gould (1986). This subset of hypothesized causal relationships, as depicted by the arrows, is presented in the descriptive model in Figure 2. This latter model contains the independent, intervening, and dependent variables investigated in this R&D project. This model is a descriptive change model and should be interpreted as such. To test the hypothesized relationships, a set of structural equations and models, following the notation of Kenny (1979), was developed. These will be discussed later in relation to specific hypotheses being tested.

This research project involved four studies conducted over a period of 13 months. Three were controlled laboratory studies focusing on rating accuracy, while the fourth study was a more methodologically based investigation. This paper first covers the general purpose and hypotheses underlying the research, then reports each study individually. Implications of the findings for the Air Force's Job Performance Measurement Project are addressed within each study.

General Background for Project

The quest for better measurement of individual job performance has generated considerable empirical research in Industrial/Organizational Psychology. However, we are still faced with the uneasy feeling, for both scientists and practitioners, that we are not "really" doing a good job in measuring job performance. Landy and Farr (1980) expressed this feeling in their review of the literature. They strongly urged researchers to stop searching for the best format as the way to improve the quality of performance ratings, and begin looking at individual differences in personal characteristics of raters or other factors that may affect rating quality.

In another sharp criticism of the ongoing performance measurement research, Hakel (1980) observed that research aimed at reducing traditional psychometric errors in performance ratings, which he relabeled "effects," was not contributing significantly to improving the quality of performance measurement practices. Subsequently, other researchers have reiterated his argument and have collected data to demonstrate that the traditional psychometric errors of halo, leniency, and range restriction may contain more than error variance (Bartlett, 1983; Hedge & Kavanagh, 1983; McIntyre, Smith, & Hassett, 1984; Wherry & Bartlett, 1982). Thus, efforts to improve the quality of performance ratings through a reduction of psychometric errors appear a somewhat illogical direction for research.

In a recent, comprehensive review of the performance appraisal literature, Kavanagh et al. (1986) presented a descriptive model detailing the many variables that could affect the quality of performance measurement (Figure 1). Unfortunately, support for many of the hypothesized relationships in the model was weak or non-existent in the literature, primarily because the authors insisted that only accuracy, or construct validity, was acceptable evidence for determining

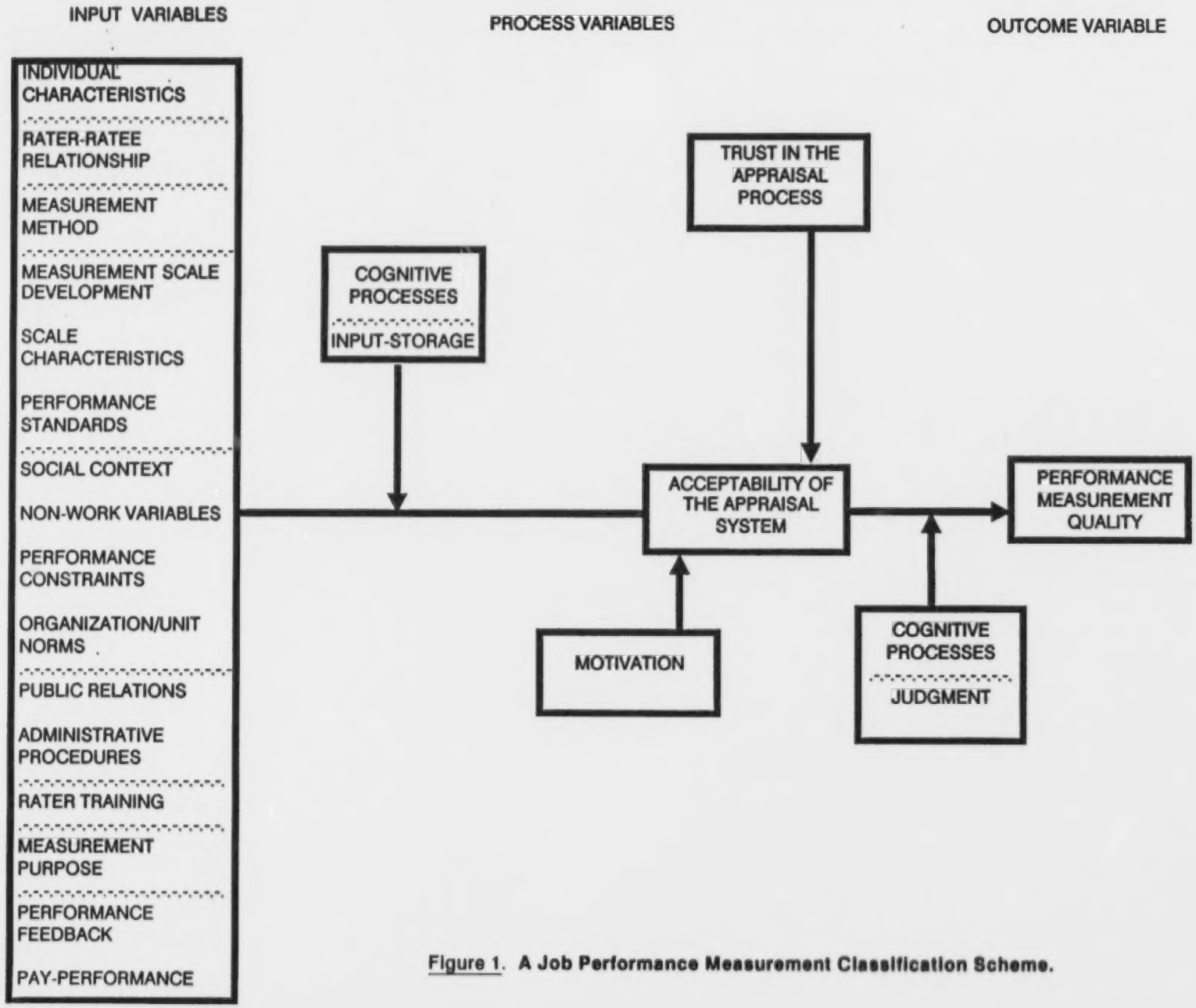

Figure 1. **A Job Performance Measurement Classification Scheme.**

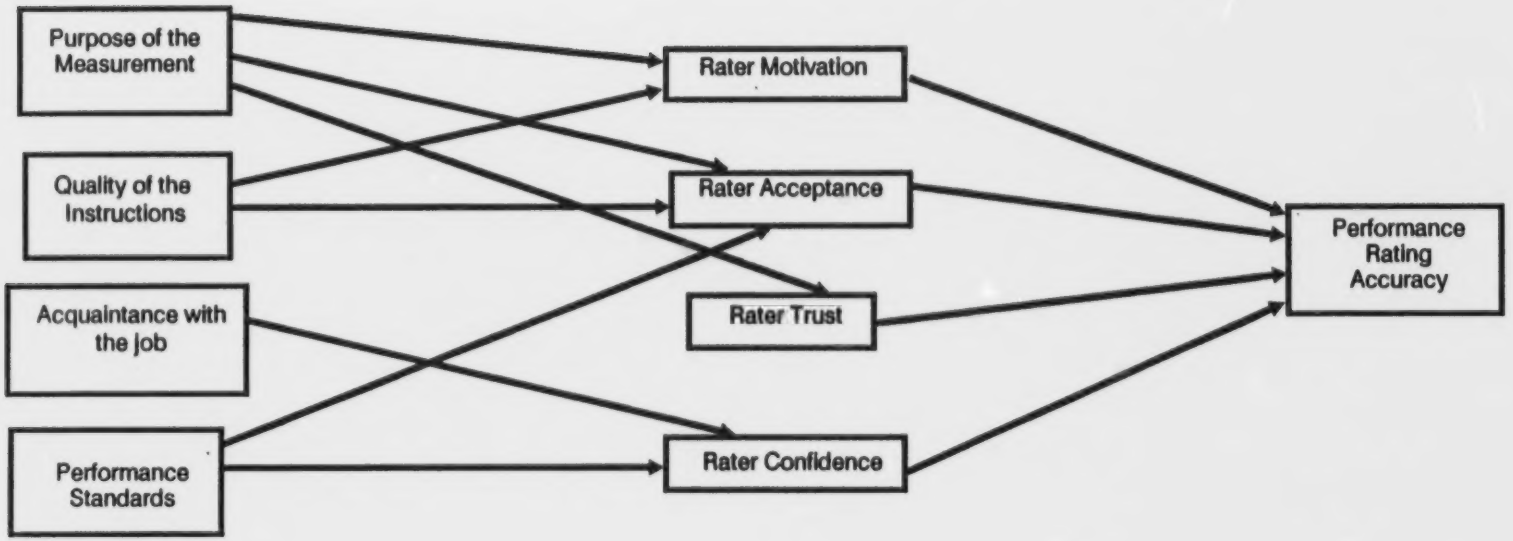

Figure 2. **Descriptive Model for Rating Accuracy Project.**

job performance measurement quality. Most of the research reviewed had used quality criteria other than accuracy. As noted by Kavanagh et al. (1986), five different criteria have been used to indicate improvements in the quality of job performance measurement: psychometric "errors," inter-rater reliability, content validity, discriminability, and construct validity or accuracy. Although the first four can be important indicators of the quality, their real value lies with the effect they have in improving the construct validity/accuracy of the measurement.

Given that accuracy is the crucial criterion against which to judge the quality of the measurement of job performance, then definitive scientific conclusions regarding the factors that affect quality of performance measurement cannot be drawn from the massive amount of literature that has relied on only one or more of the four other criteria. Therefore, the appropriate research method to test the effects of personal and organizational variables on performance measurement quality, as depicted in Figures 1 and 2, would use accuracy as the dependent variable. This logic is consistent with current theory in measurement (Nunnally, 1978) and performance ratings (Wherry & Bartlett, 1982), and has guided this R&D project.

Although concern with construct validity/accuracy has been a part of the measurement literature for some time, it took the work of Borman, Hough, and Dunnette (1976) to provide an experimental methodology to assess accuracy in performance ratings. By creating videotapes of eight different sequences of actors/employees performing a job, they were able to develop "true scores" for several dimensions of job performance. In this way, the videotapes represented a fixed, or standard, stimulus for which the true performance scores were known. Using this methodology,the performance ratings given by subjects in an experiment could be compared to the true scores, allowing one to determine how accurate the subjects were in their ratings of the actors in the videotapes. The effects of various independent variables on the accuracy of the performance ratings could now be studied. For example, Hedge and Kavanagh (1983) used videotaped performances to study the effects of different rater training programs on rater accuracy. This method also would allow determining how performance appraisal systems can be changed to improve accuracy. This was the general purpose of this research project.

Another viewpoint that has guided this research is the practicality and applicability of research results for guiding the development of a job performance measurement system (JPMS) that can be used by the Air Force to validate the Armed Services Vocational Aptitude Battery (ASVAB)--the test used by the armed services to determine qualifications for enlistment and placement within a specific job or occupational area (Department of Defense, 1984). In this case, the independent variables chosen for study are ones over which there is some degree of control. For example, personality of the rater may be found to affect rating accuracy by an organization, but most personality variables are difficult to change. On the other hand, different instructions to raters may have differential effects on the accuracy of the ratings. Instructions are reasonably controllable and thus worthy of research within the context of the JPMS. In this way, the practicality of the research to the organization helped guide what was included in this and other JPMS research projects. As noted by Banks and Murphy (1985), considering organizational constraints while planning and conducting research helps to narrow the "research-practice gap in performance appraisal."

Research Variables

Four independent variables were used in this research project. The first, purpose of measurement, concerns the use of the performance ratings. In this project, the "purpose" variable was operationalized in terms of whether the performance ratings were being collected for "operational" purposes (e.g., a promotion decision) or "for research purposes only."

The second independent variable, acquaintance with the job, refers to the amount of experience the rater has with the job being performed. In this project, biographical questionnaires were used to identify raters having varying levels of experience with the target tasks.

The third independent variable, performance standards, concerns whether or not specific anchors exist on the performance rating scales. These performance standards are meant to be much more detailed than a typical Behaviorally Anchored Rating Scale (BARS). In the routinely accepted method for creating a BARS following Smith and Kendall (1963), highly specific performance items frequently are eliminated during the retranslation procedure. This results in performance dimension descriptions that are more general in nature and have anchors that do not adequately define the performance standard for the job dimensions. Performance standards for the rating forms developed by Borman (1978) were developed using Subject-Matter Experts (SMEs) in contrast to the BARS format (Smith & Kendall, 1963).

The fourth independent variable, quality of instructions, refers to the amount of detail and clarity in the instructions accompanying the performance rating form. Quality may also be dependent on the mode of presentation. In this project, "quality of instructions" was operationalized by level of detail and three modes of presentation.

Four intervening variables are depicted in Figure 2. For this project, rater motivation is conceived as the internal drive to make an accurate rating. As such, it may be the most global construct of the four intervening variables. Rater acceptance is defined as the rater's willingness to complete the performance ratings because the ratings are seen as worthwhile for the organization or research study. Rater trust relates to the trust the rater has that the performance ratings will be used for their stated purpose. This concept may also encompass the rater's trust that other raters will "play fair" with the performance appraisal system. The final intervening variable being considered here is rater confidence, the degree to which the rater believes he/she can accurately reflect the ratee's performance on the appraisal form. This type of confidence is based on the rater's perceived ability to distinguish good from poor performers using the performance rating form.

Literature Review

Purpose of Measurement

As noted by Kavanagh et al. (1986), differing purposes of the performance measurement will create different contexts that can impact on the quality of the measurement. The purpose of the measurement can create differing demands on raters (Wherry & Bartlett, 1982), and may lead to "motivated errors" (Kane, 1980) that can seriously impact on rating accuracy. Most empirical studies examining this issue have contrasted performance ratings being used for administrative purposes (pay raise or promotion) with ratings collected for use either in research or for the development of the individual employee.

The first research studies on this topic were done in military settings (Berkshire & Highland, 1953; Taylor & Wherry, 1951). When the purpose of the performance ratings was administrative versus research, Taylor and Wherry (1951) found significantly more favorable (i.e., more lenient) ratings were given. Berkshire and Highland (1953) did not find this effect. In a different setting, Bernardin, Orban, and Carlyle (1981) found performance ratings given to rookie police officers were significantly more favorable when the ratings were going to be used for administrative

purposes as opposed to their use as feedback to the officers. The studies, however, did not use rating accuracy as the dependent variable.

There have been a number of other studies examining the perceived purpose of the ratings in the context of students' evaluations of college instructors. All of these studies found that students' ratings of their instructors were higher when students were told the ratings were being used for personnel or administrative decisions versus other purposes (Aleamoni & Hexner, 1973, 1980; Centra, 1976; Driscoll & Goodwin, 1979; Sharon & Bartlett, 1969; Smith, Hassett, & McIntyre, 1982). Again, these investigations used leniency error, and not rating accuracy, as the primary index of rating quality.

Two fairly recent studies (McIntyre et al., 1984; Zedeck & Cascio, 1982) examined purpose of measurement as a factor, along with different rater training programs, using both the traditional psychometric indices and rating accuracy as indicators of rating quality. However, the results of these studies were contradictory. McIntyre et al. (1984) found purpose of measurement had no effect on accuracy, but rater training programs did; Zedeck and Cascio (1982) found no effects of training, but significant effects on accuracy due to measurement purposes. Although McIntyre et al. (1984) discussed reasons why such differences between the studies may have occurred, the important point for this research is that the role of measurement purpose in rating accuracy issue has not been empirically resolved.

One of the main efforts of the study covered in this paper was to examine the effects of two purposes of performance ratings, promotion decisions versus research only, on the quality of measurement to include both traditional psychometric and accuracy criteria. Since previous research results have been contradictory or deficient in that accuracy criteria were not used, it was hoped that this research would provide some insight to help explain these previous results. Furthermore, it should help to indicate how performance rating data should be collected within the JPMS currently ongoing with the Air Force Human Resources Laboratory. A detailed explanation of the hypothesis regarding the purpose of measurement variable in relation to Figure 2 is contained in the "Research Hypotheses" section of this paper.

Acquaintance with the Job

The second dependent variable in this research, acquaintance with the job, has received little attention in the performance measurement literature. Although it appears almost axiomatic that a rater more acquainted with a job would provide a better, more accurate rating of an employee in that job than would a rater with less acquaintance, no direct evidence of this relationship exists.

There have been several studies that have examined various characteristics of the relationship between the rater and the ratee. The degree of responsibility the rater had over the ratee's previous performance (Bazerman, Beekun, & Schoorman, 1982), the rater's familiarity with the ratee's previous performance (Jackson & Zedeck, 1982; Scott & Hamner, 1975), and the degree of acquaintance between the rater and ratee (Freeberg, 1968) have all been shown to affect the quality of job performance measurement. The degree of acquaintance variable is most interesting. The rater must be somewhat acquainted with the ratee's performance to complete the performance ratings. In fact, most authors argue that the rater must have had the opportunity to observe job-relevant behaviors or else the rating will contain error (Borman, 1974). Stone (1970) has argued that as the degree of acquaintance increases, the possibility of bias in terms of halo increases, particularly if the rater and ratee become friends. This logic is consistent

with Corollary 3a and Theorem 4 of Wherry's theory of rating (Wherry & Bartlett, 1982); however, it has not been directly tested in the performance measurement domain.

This degree of acquaintance variable has, however, two dimensions. The rater can differ in the degree of acquaintance he/she has with the task requirements of the job, and the degree of acquaintance with the employee doing this job. The latter meaning of acquaintance has been the focus of the research discussed above; however, it was not examined in the present research. Although this may be a potentially powerful variable in terms of its effect on rating accuracy, it was felt that the former meaning of acquaintance with the job (knowledge of the task requirements) was more important, for both research and practical reasons, for this research. This variable has simply not been investigated in the empirical literature, although it has been generally assumed that a rater must be acquainted with the job before an accurate appraisal of a person doing that job can be done. In terms of the JPMS effort, it is important to determine what degree of acquaintance with a job is necessary to provide accurate performance ratings, in order to determine what raters are appropriate for JPMS.

Performance Standards

Performance standards that provide more specific anchors for job performance rating scales were first employed by Kavanagh, Hedge, DeBiasi, Miller, and Jones (1983) in the development of a new performance appraisal system for a hospital corporation. After management expressed their extreme disapproval of a rating format derived using the standard BARS technique (Smith & Kendall, 1963), a Behaviorally Anchored Summary Scale (BASS) was developed using specific performance standards judged (by consensus) acceptable to management. Thus, the definition of what constituted each standard was decided by the management of the organization, not by the industrial psychologist through statistical decision rules only. Adding this step to the BARS technique assured that the performance rating form reflected the mores, climate, and culture of the organization in which the form was embedded.

This need for the use of performance standards on a rating scale was identified in a review of legal cases regarding compliance with Equal Employment Opportunity Commission (EEOC) guidelines on the use of performance appraisal in personnel decisions (Cascio & Bernardin, 1981). These authors argued that the performance appraisal form must have performance standards if it is to be in compliance with legal decisions and the EEOC guidelines. If the use of performance standards can also improve the accuracy of the measurement, then this practice would be doubly rewarding. Although there are arguments for the use of performance standards (Alewine, 1982; Kirby, 1981; Morano, 1979), no empirical evidence exists to support their use. The use of performance standards on the rating scale was tested for the first time in this project.

Quality of Instructions

As noted earlier, this variable includes both the level of detail and clarity of the instructions that accompany the rating scale and the way the rating task is presented to the raters. Although we could identify no research addressing these variables within the job performance rating literature, they are extremely important to the JPMS project of AFHRL. Since the performance measurement system resulting from the JPMS project is intended, in part, to be used to validate the ASVAB, there are significant practical issues regarding the large-scale data collection effort needed to complete this validation project. Perhaps the single most significant issue is how to

collect these job performance data in the most accurate and cost-effective manner. Thus, the detail of instructions and mode of presentation variables were evaluated in this research project.

Methodological Issues

Two methodological issues were also addressed in this research project. The first issue deals with the technique and stimulus materials used to conduct research on performance rating accuracy. The Borman et al. (1976) method described earlier uses videotapes as the standard stimuli on which accuracy of raters' judgments is determined. Another technique uses "paper-people," or performance vignettes, to examine the relationship between independent variables and rating accuracy in the performance appraisal literature. The vignette approach uses narrative descriptions of employees performing a job at varying performance levels. The true score matrix is determined either by specification of specific "target scores" in the script writing process or, in a few cases, by expert judges who rate the vignettes.

An important methodological and empirical issue to be resolved is whether the verification (or non-verification) of hypothesized relationships between independent and dependent variables in rating accuracy research depends on which true score technique is used to study the relationships. For example, in testing the empirical relationships depicted in Figure 2, does it matter whether one uses the videotape or the vignette method? If it were found that the purpose of measurement had a differential effect on measurement quality depending on whether the videotape or the vignette method were used, what could be concluded? This concern is closely linked to the JPMS project since the "best" true score technique must be established if specific, prescriptive advice regarding the design of a performance measurement system for use in validating the ASVAB is to stand the test of close scrutiny. Thus, these two different approaches to the study of rating accuracy were carefully evaluated in this project.

Before leaving this issue, it is important to note that the contradictory results found for the purpose of measurement in two earlier studies (McIntyre et al., 1984; Zedeck & Cascio, 1982) involved two different research methods. As noted by McIntyre et al. (1984), their study used the videotape method while the other study used the vignette approach. Without belaboring this point, these different methods requiring different capabilities of the raters may have been the main reason for the differing results.

The second methodological issue addressed in this research project involves the validity of the "true score" matrix developed for use with the Borman videotapes. This set of scores was developed in 1975 using "expert" judges. These judges were industrial psychologists who provided ratings of the performance of the individuals shown in the videotapes. The empirical and methodological question is whether another group of SMEs with different backgrounds and training would derive the same or a different set of true scores. If these SMEs provide different, and better, true scores, the implications for the JPMS project are clear. This "SME-derived" set of target scores should be used in evaluating the hypothesized relationships in Figure 2. Of course, the issue of which set of true scores is best is controversial. Central to this controversy is the definition of who are the "best" SMEs to provide true scores. It can be argued that the SMEs selected for this research project are better than those used to establish the original true score matrix for the Borman tapes. However, whether the SMEs used in this project are the "best" is a philosophical argument which would be very difficult to resolve empirically.

Intervening Variables

This research was also concerned with examining the role of the intervening variables depicted in Figure 2 of rater motivation, acceptance, trust, and confidence. It is assumed that these variables moderate, to some degree, the effects of the independent variables on performance rating accuracy. Previous research on performance rating accuracy has ignored these types of personal variables; however, examination of the role of these variables in terms of their impact on rating accuracy is both consistent with recent emphasis on cognitive variables in the performance appraisal process (Feldman, 1981; Landy & Farr, 1980) and with the practical need to understand the raters' motivation in the JPMS project.

Research Hypotheses

To facilitate an understanding of the hypothesized relationship derived from Figure 2, structural models were constructed. To accomplish this, the notation established by Kenny (1979) was used. Further, signed relationships corresponding to specific hypotheses were indicated on these structural models. It should be noted that the use of signed relationships is an extension of the standard symbols used in structural modeling, which typically contains only the hypothesized causal relationships without positive or negative signs. However, this was an excellent way to develop the hypotheses of this project for expository purposes. These models with signed relationships for the causal parameters were extremely helpful in establishing a priori statements of the hypotheses of this research, which, according to Kenny (1979), is a very critical step in social science research.

For the purpose of the structural models, the variables of interest for this research contained in Figure 2 have been assigned the following notation: purpose of measurement, X1; quality of instructions, X2; acquaintance with job, X3; performance standards, X4; rater motivation, Z1; rater acceptance, Z2; rater trust, Z3; rater confidence, Z4; performance rating accuracy, Y. The letters U and V represent residual disturbance terms that include all unspecified causes of the intervening or dependent variables. The lowercase letters in the structural models are the causal parameters, and their interpretation is straightforward.

The structural model and equations for the purpose of measurement variable are contained in Figure 3. As can be seen, this independent variable is hypothesized to directly affect rating accuracy, and its effect is represented by the causal parameter a. Likewise, the independent variable is hypothesized to affect three intervening variables: rater motivation, rater acceptance, and rater trust. The intervening variables are also hypothesized to affect the dependent variable. The disturbance terms, U and V, represent all of the unspecified causes for the changes in the dependent and intervening variables, respectively.

Figure 3a depicts the hypotheses regarding the independent and intervening variables when the purpose of the performance measurement is for administrative or operational use, such as a promotion decision. The negative and positive signs attached to the causal parameters indicate the hypothesized direction of the effects for the independent and intervening variables, and are based on the literature cited earlier. Thus, it is hypothesized that, when the purpose of performance measurement is for administrative use, there will be a negative effect on accuracy, and, most likely, an increase in leniency. It is further hypothesized that this performance measurement purpose condition will negatively affect two of the intervening variables, rater acceptance and trust, while positively affecting rater motivation. These hypothesized effects will be carried through to affect measurement quality as indicated by e, f, and g in Figure 3a.

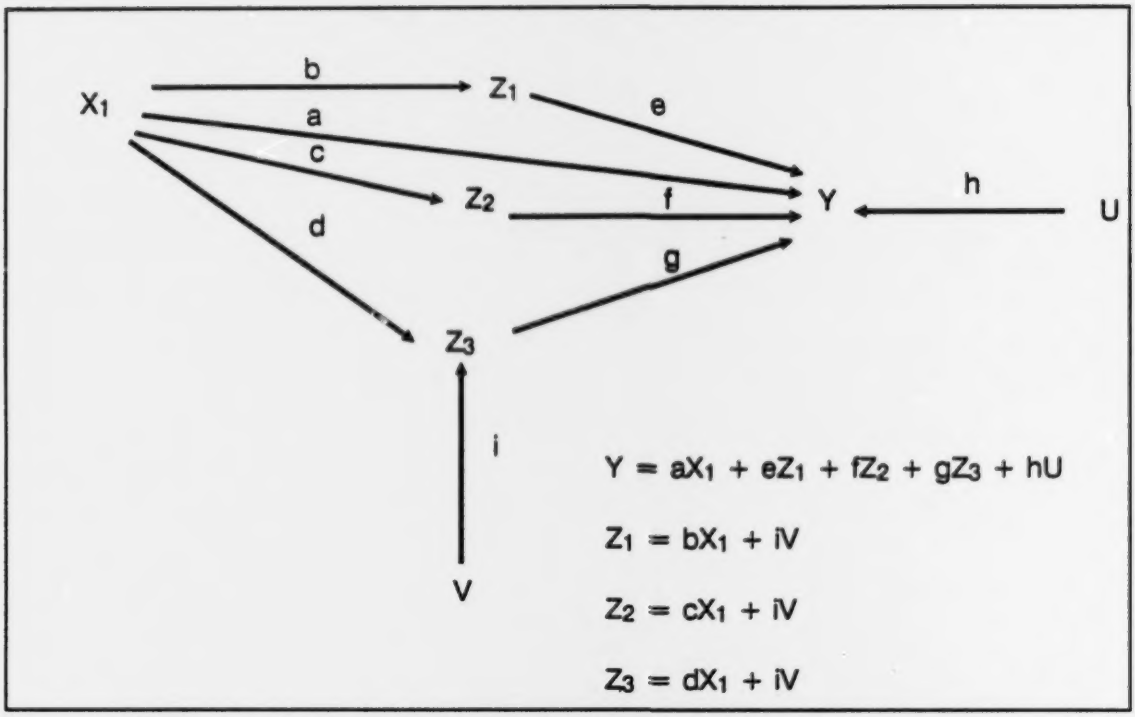

Figure 3. **Structural Model and Equations: Purpose of Measurement.**

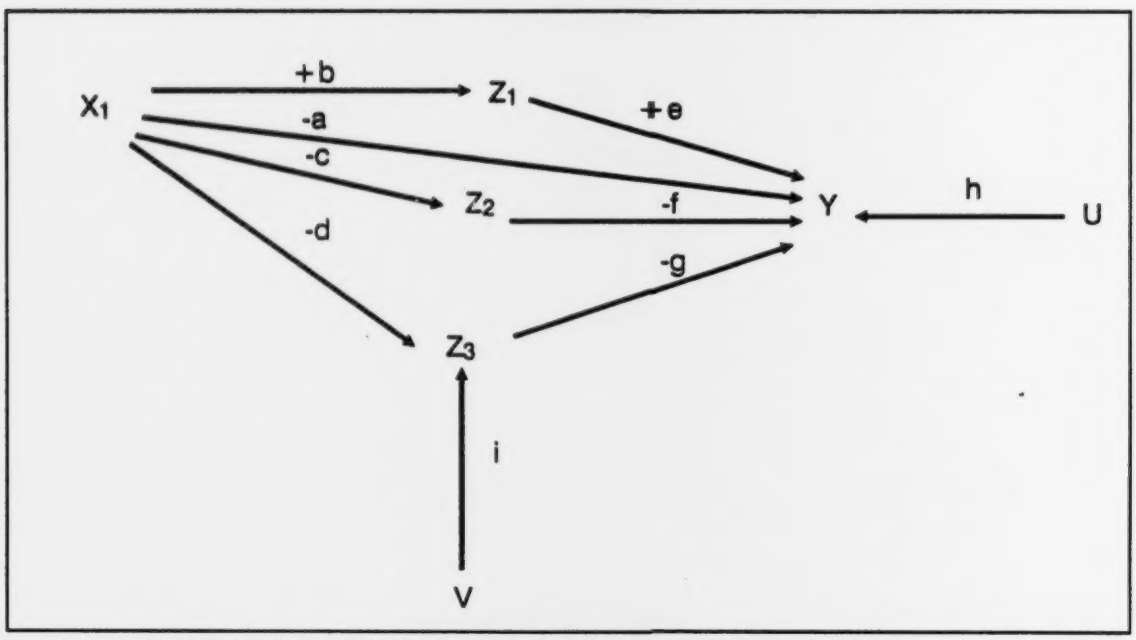

Figure 3a. **Model with Signed Relationships: Operational Purpose.**

Figure 3b depicts the hypothesized relationships among the research variables when the purpose of the performance measurement is for validation research. In contrast with Figure 3a, it is hypothesized that, in general, the measurement quality will be better, both in terms of the main effect of this condition and the impact on the intervening variables. Note, however, the negative relationship hypothesized between the independent variable and rater trust, as well as the negative relationship hypothesized between rater trust and the dependent variable.

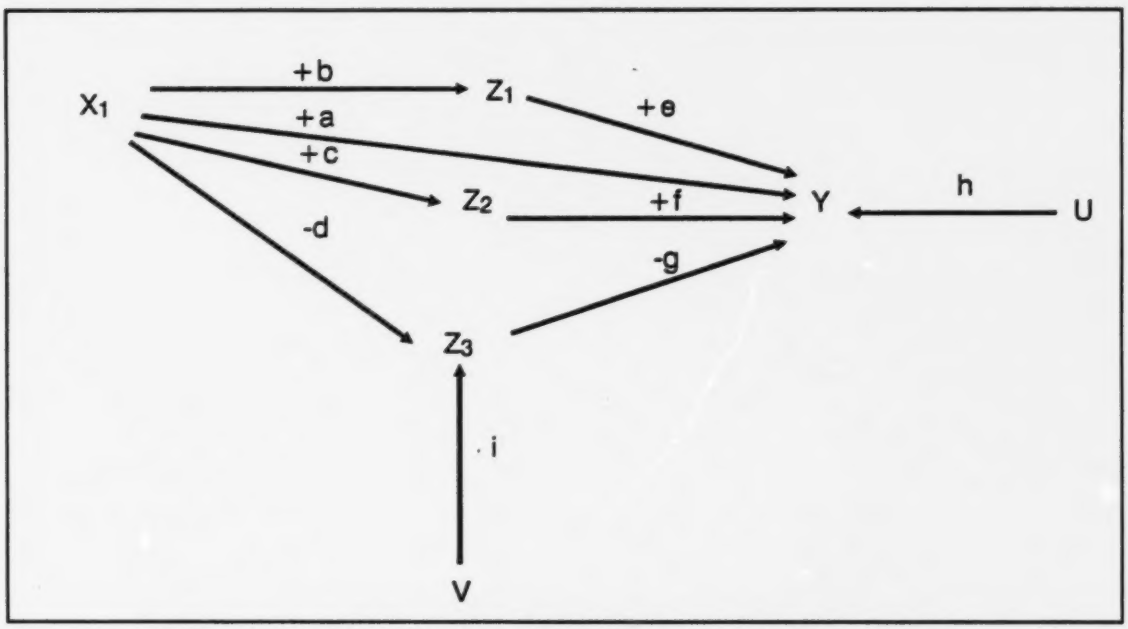

Figure 3b. **Model with Signed Relationships: Validation Purpose.**

The hypotheses regarding acquaintance with the job are contained in Figures 4, or 4a, and 4b; and the symbols are to be interpreted as was done in the previous figures. Based on common sense and the sparse literature available, it is hypothesized that the rater's acquaintance with the job on which the ratee is being evaluated will affect both rater confidence and rating accuracy; i.e., the higher the degree of acquaintance with the job, the higher the confidence and the more accurate the ratings.

The hypotheses regarding the difference between a rating scale format based only on BARS technology versus one with the addition of performance standards are depicted in terms of structural equations and models in Figures 5, 5a, and 5b. Examination of the signed relationships indicates that the rating form with performance standards is hypothesized to be superior to the form with BARS in terms of its effect on rating accuracy. It should also be noted that this is due to the differential effects of the two conditions on the intervening variables, as seen in the figures.

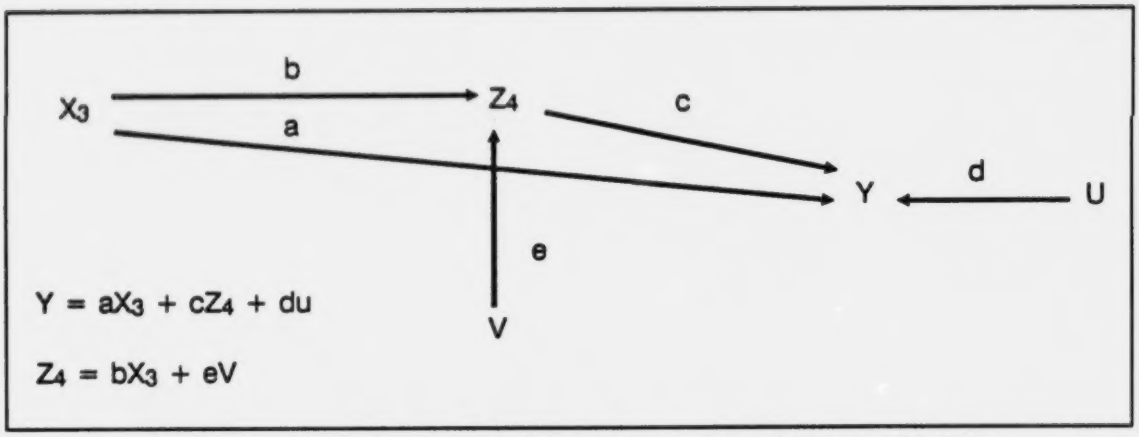

Figure 4. Structural Model and Equations: Acquaintance with Job.

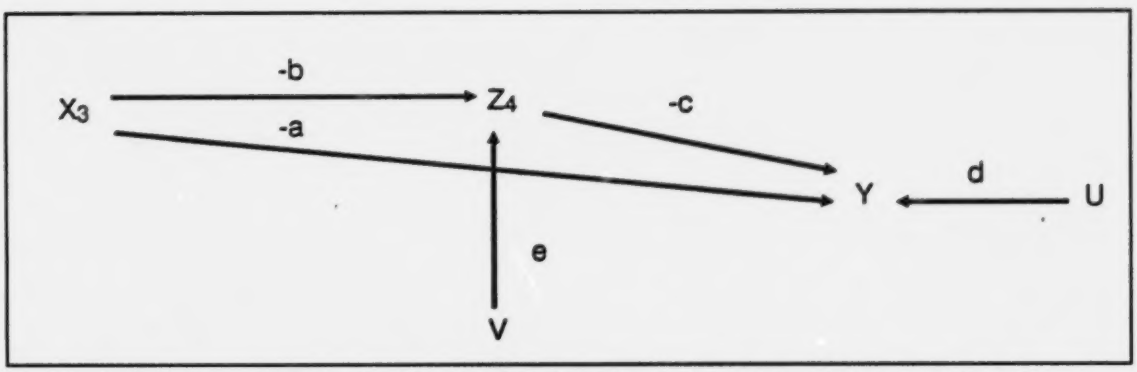

Figure 4a. Model with Signed Relationships: Low Acquaintance with Job.

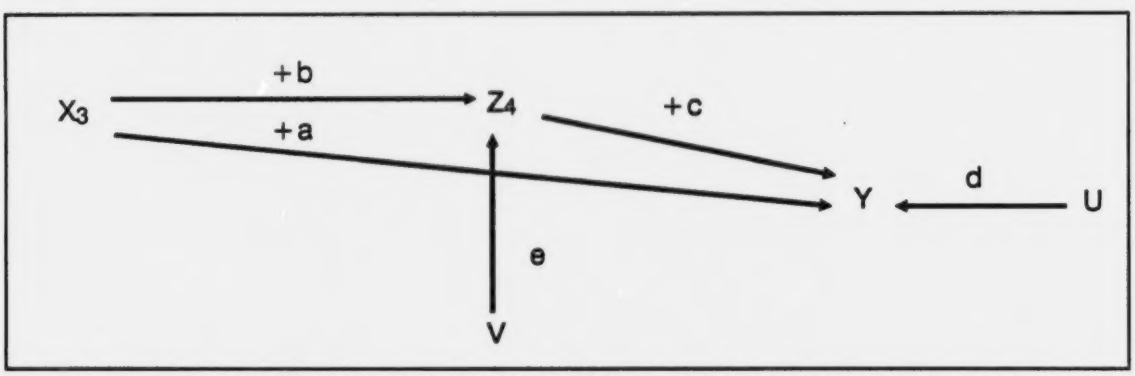

Figure 4b. Model with Signed Relationships: High Acquaintance with Job.

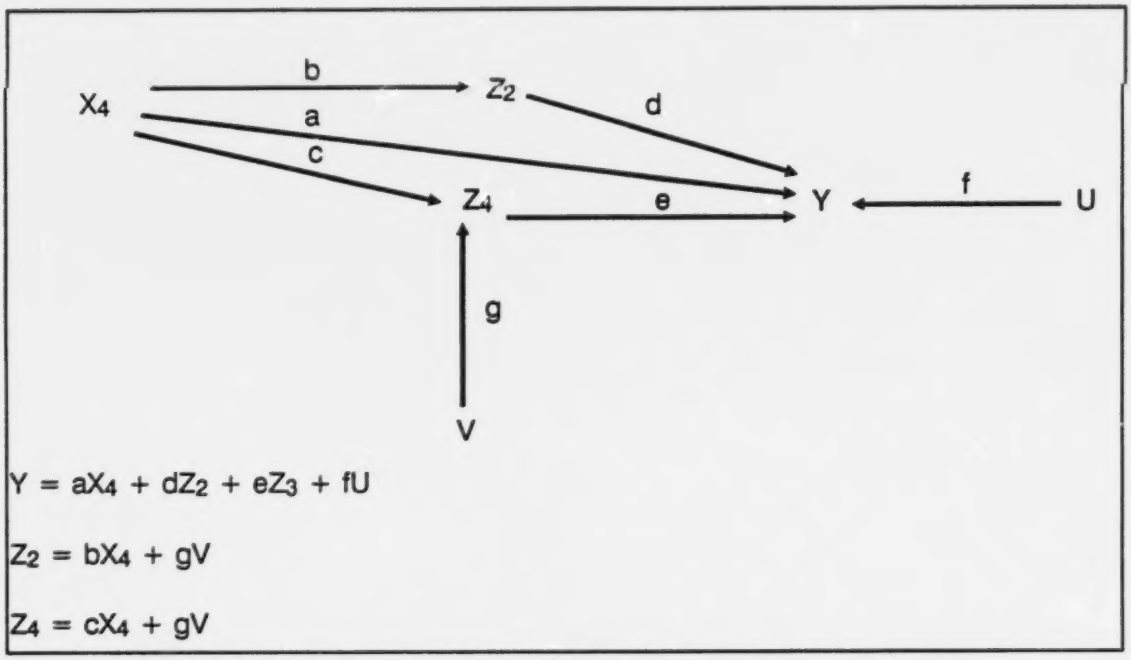

Figure 5. **Structural Model and Equations: Peformance Standards.**

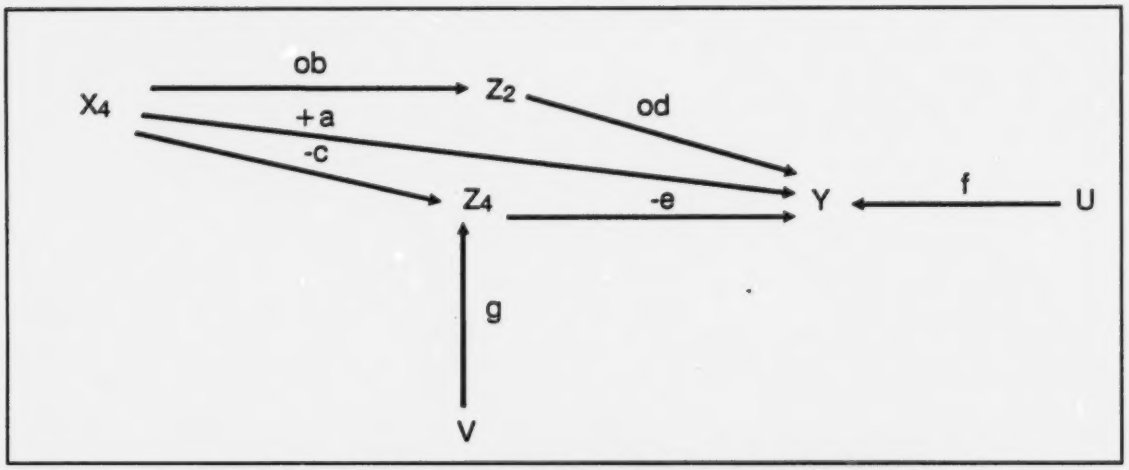

Figure 5a. **Model with Signed Relationships: BARS Format.**

Hypotheses concerning the quality of instructions and the amount of detail are contained in Figures 6, 6a, 6b, and 6c. Since there is no empirical literature on the mode of presentation with regard to collecting performance ratings, the a priori hypotheses represent exploratory, common sense ideas.

As can be seen in the figures for this variable, it is hypothesized that the amount of detail on the rating scale will impact on the accuracy of the ratings. For this project, there were

three levels of detail: small (or standard), moderate, and large. Comparison of Figures 6a, 6b, and 6c indicates that increasing the amounts of detail in written instructions is hypothesized to have positive effects on both the intervening variables and rating accuracy.

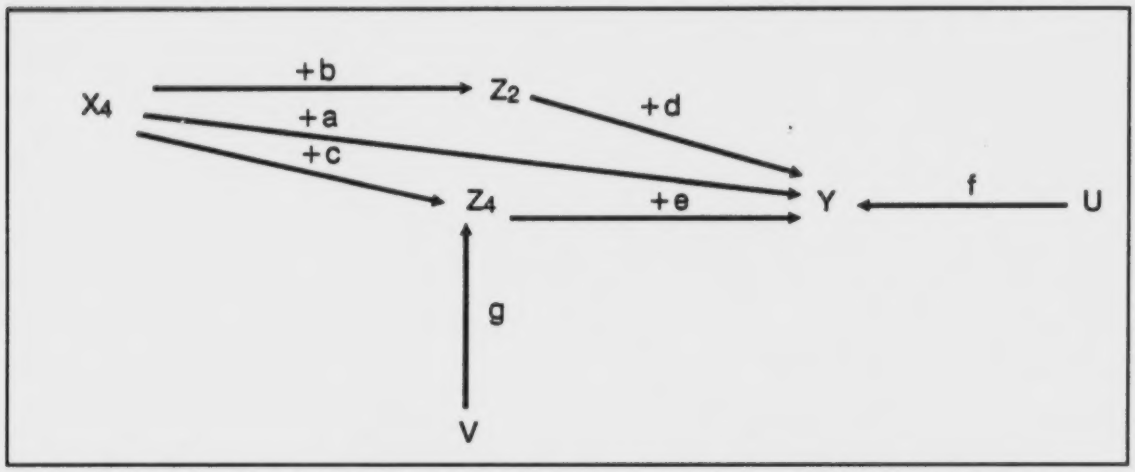

Figure 5b. **Model with Signed Relationships: Performance Standards Format.**

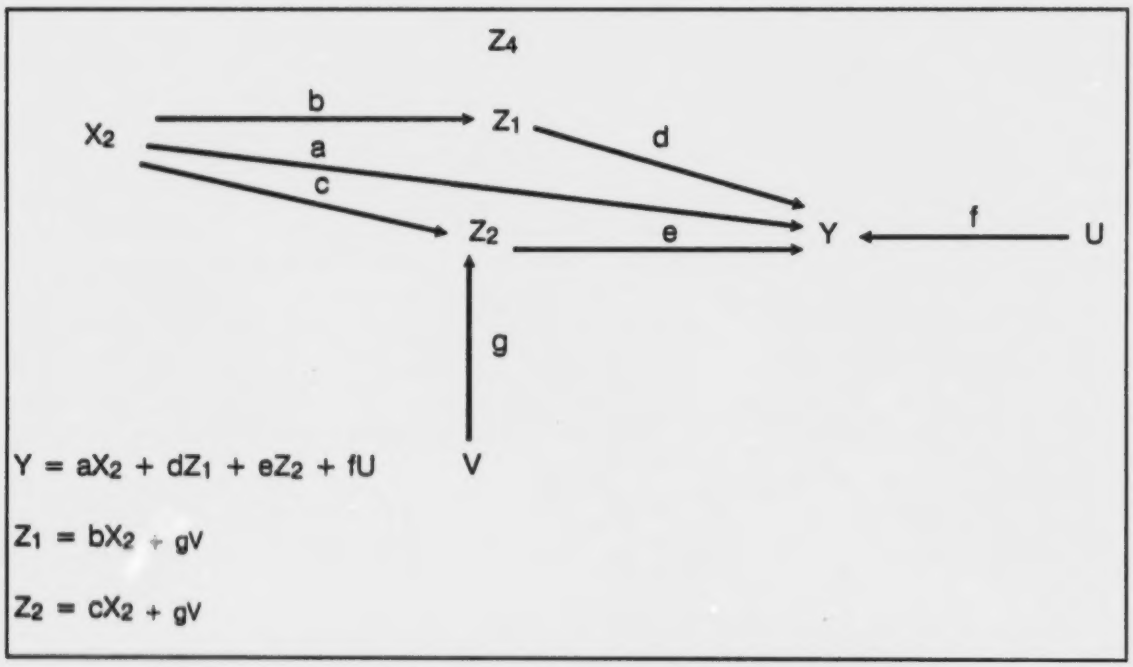

Figure 6. **Structural Model and Equations: Quality of Instructions.**

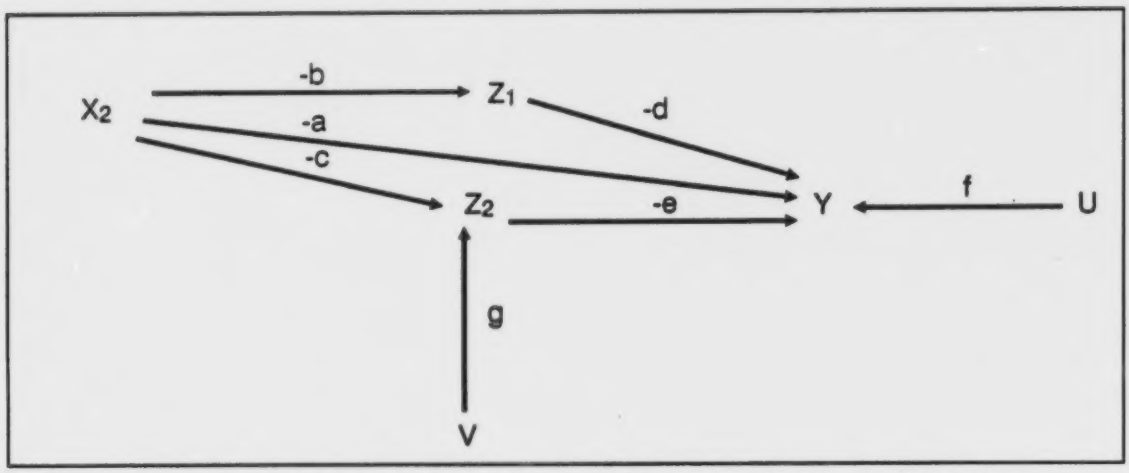

Figure 6a. **Model with Signed Relationships: Small Amount of Detail.**

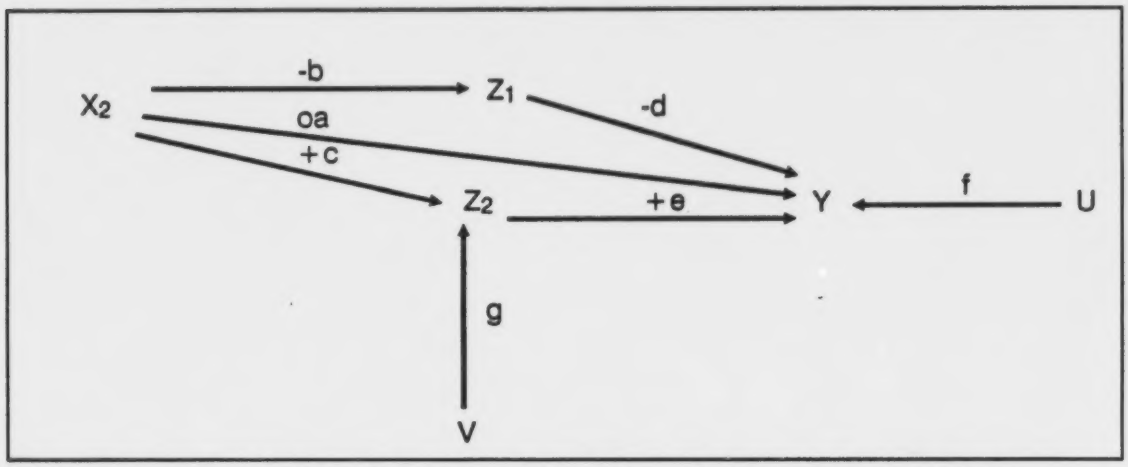

Figure 6b. **Model with Signed Relationships: Moderate Amount of Detail.**

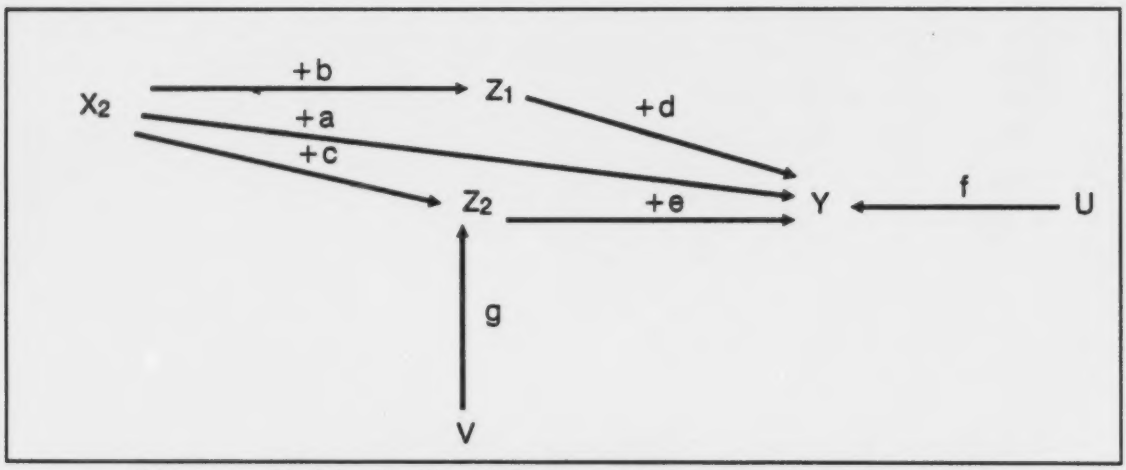

Figure 6c. **Model with Signed Relationships: Large Amount of Detail.**

With regard to a comparison between the videotape and vignette methodology, it is generally hypothesized that the videotape technique will be superior because of the significantly better sensory information it provides as contrasted with that provided by vignettes of employee performance. By analogy, this may be likened to the contrast between the informational content gained by reading a script and by actually seeing the play. Since it is hypothesized that all relationships among the research variables in the videotape condition will be more positive than in the vignette condition, it seemed unnecessary to draw the structural models.

Finally, the research using the new group of SMEs is anticipated to produce a new set of true scores superior to those created by Borman (1978). SMEs with specific background and training in personnel, as well as experience with the performance appraisal feedback interview, were chosen to participate in this study. These SMEs were also used to develop the "performance standards" rating scale format for this research, and to discuss the possible "criterion deficiency" of both the videotapes and the rating scale used by Borman (1978).

II. STUDY 1

This first study was concerned with testing the research method and the following independent variables: purpose of measurement and rater acquaintance with the job.

Method

Experimental Design

Based on the hypotheses of this research, a completely randomized, 3 X 2 factorial, fixed effects design was used to collect the data. This allowed for three levels of the first factor, experimental method, and two levels of the second factor, purpose of the performance measurement.

The first factor, experimental method, had three levels to reflect fully the problems with these different accuracy paradigms as described in the literature. The first two conditions for this factor are the ones that normally come to mind, i.e., a written vignette versus a videotape of the same job performance sequence. However, close examination of the literature describing these two techniques revealed an important methodological distinction, not noted before, between the two. In the videotape technique, the raters watch a tape of the performance of the job incumbent and then are asked to rate this performance without an opportunity to review the tape while doing their performance ratings. In the vignette technique, raters read the written material, and are allowed to refer back to it while completing their performance ratings. Therefore, in order to have adequate comparison data, an additional experimental condition was used in this study. The first condition was the videotape, with no opportunity to refer back to the tape. The second condition was the vignette with an opportunity to refer back to the written condition, and the third was a vignette with no opportunity to refer back to the written material.

The second factor, purpose of the measurement, had two levels--administrative versus research. As will be seen in the description of the experimental procedures, the administrative purpose was created by informing the raters that the persons they were rating, in either the

videotapes or the vignettes, were being considered for promotion and that the raters' ratings would be used in the promotion decision. In the research condition, the raters were told that their ratings were being used in a research study to validate a set of tests and exercises used in a managerial assessment center.

Subjects

Data were collected from 134 graduate students in both the evening and full-time Master of Business Administration (MBA) program in the School of Business at the State University of New York at Albany (SUNYA). Although only 90 raters were necessary for sufficient power, given the experimental design (Cohen & Cohen, 1975), the additional raters were necessary to empirically establish the reliabilities of the measures of the intervening variables. For the analyses testing the hypothesized main effects and interactions, Multivariate Analysis of Variance (MANOVA) and Analysis of Variance (ANOVA), it was necessary to randomly eliminate raters from some cells to achieve equal cell numbers such that the expected mean squares could be correctly estimated. This resulted in 18 raters per cell, more than sufficient for the power analysis (Cohen & Cohen, 1975). Thus, the results to be discussed in subsequent sections have varying numbers of raters, reflective of the varying investigations within this study.

Stimulus Material

The two sets of videotapes with the original scripts used to create them (Borman et al., 1976) were made available for this research project (Borman, personal communication, 1984). There are two sets of tapes, each with eight different persons performing the job to be rated. "True scores," using expert raters, for all of these tapes were developed as part of the Borman et al. (1976) original work. One set of videotapes shows the interaction between a college recruiter, from the Personnel Department of an engineering firm, and a college senior. The second set of tapes involves a performance appraisal interview between a supervisor and subordinate manager in an engineering firm.

After careful examination of the videotapes from the Borman et al. (1976) study, it was decided that only the performance appraisal tapes would be used since the recruiter videotapes were deemed out-of-date. Further, it was decided, in order to save time, to use less than all eight job performance sequences. This decision was based on research that indicated that five sequences produced reliable estimates of the raters' accuracy (Bernardin, personal communication, 1984). Since six of the videotaped sequences were deemed technically superior in terms of video and audio presentation, these six performance sequences were used for this study. These tapes and scripts, described in Borman et al. (1976), were the ones used for all the research studies in this project.

Since both the original scripts and videotapes were available for the six different sequences of manager performance in the performance appraisal interview, it was decided to use the scripts as the vignettes in the two vignette conditions of this research. It was felt, for comparison purposes, these scripts were the best available "paper people" descriptions that represented the performance depicted in the videotapes.

Research Variables

Independent Variables. The first independent variable, purpose of measurement, was manipulated through instructions to the raters that ratings were for a study funded by a major organization to evaluate the managers on the videotapes (or scripts) either for potential promotion or for use in research involving the validation of a managerial assessment center. In both purpose of measurement conditions, the importance of the study was emphasized, as an attempt to control the importance variable.

This was based on an examination of the previous literature where it is apparent that the "importance of the ratings" has not been controlled. In previous research comparing the purpose of measurement, little attention has been paid to the unintended social forces in laboratory research (see Duffy & Kavanagh, 1983). In experimental research on performance accuracy with a purpose manipulation, no attention has been paid to the social forces caused by the manipulation of the importance of the ratings. Thus, in a typical study comparing purpose of measurement, one would expect that performance measures collected for personnel or administrative purposes would be seen as generally more important than those collected for research or employee growth purposes. Thus, the importance aspect was controlled in this study through the use of scripts for the experimenters that emphasized the importance of the performance ratings several times.

The second independent variable, "acquaintance with the job," was assessed by questionnaire at the beginning of the academic semester. All MBA students completed a biographical survey on the first day of classes to assess their types and levels of experience. As part of this questionnaire, several items were included to assess the acquaintance of these students with the job of a supervisor or engineering manager and their experience regarding performance appraisals and feedback interviews. Thus, the questionnaire contained the following items designed to assess this acquaintance with job variable:

Total years of full-time work experience?

Total years of experience as a supervisor/manager?

If you have completed performance appraisals for employees under your supervision, what is the approximate number you have done to date?

If you have provided feedback interviews on employees' performance, what is the approximate number to date?

Have you ever been a supervisor for engineers? If so, for how many years?

It was felt that these questions would differentiate among those raters who had varying degrees of acquaintance with the job. This questionnaire is included in Appendix A.

The third independent variable, the experimental technique, had three different conditions. The first condition was created by using the six videotape sequences. The written scripts of the videotapes, with instructions to not refer back to the scripts, created the second condition; and the written scripts, with instructions that allowed the raters to refer back to the scripts when making ratings, was the third condition.

Intervening Variables. The intervening variables of rater motivation, acceptance, trust, and confidence have not been studied previously in the way in which they have been described in this study. As a result, it was necessary to conduct a thorough psychometric development (Nunnally, 1978) for these variables. This involved: (a) operational definition of the four constructs (see the description in the Introduction); (b) hypotheses regarding the existence and operation of these variables in regard to the investigation of interest (see the hypotheses of this study); (c) generation of the item pool; (d) semantic calibration of the item pool; and (e) empirical verification. The empirical verification of the measurement of these variables will be discussed with the results of this study.

Dependent Variables. The dependent variable for this research is the quality of the performance ratings made by the raters. Consistent with previous research (McIntyre et al., 1984), estimates of distance accuracy, correlational accuracy, halo, leniency, and range restriction were computed. Due to the questions raised regarding these two methods, videotape versus vignette, a measure of "confidence in the specific rating" was collected from the raters. This was done by having the raters rate, on a 5-point scale, how confident they were with their ratings of each videotape or written vignette. These confidence ratings were useful in explaining the hypotheses of this study, and served as an additional dependent variable.

Procedure

Subjects (raters) were randomly assigned to one of the six treatment conditions in the design. The data were collected as part of a class session on performance appraisal in three sections of a graduate course in Human Resources Management. The raters were initially briefed on the general purpose of the research, the importance of the data collection, and their role. No experimental conditions were introduced at this time except the importance variable. All raters were told that the study was a "$100,000 project awarded to SUNY-Albany to rate the performance of managers in a performance appraisal interview situation." After this brief introduction, raters were asked to complete an "informed consent form," which all did.

The raters were then randomly split into two equal groups, and one of these groups went to another classroom. The two groups were split for the purpose manipulation, with one group told that the study was for administrative purposes and the other, research purposes. Since there were three sections of the class, it was determined, by random selection, to put each of the sections in either the videotape (VT), the script refer back (SRB), or the no refer back script (NRBS) condition.

In the administrative purpose condition (ADMIN), the subjects were told that the rating data that they were providing on the "real" managers in the performance sequences were going to be used to help determine which of six managers would be promoted to the next higher level of management. It was stressed that the ratings were a significant piece of the total information that would be used to make the promotion decision, and that the sponsoring organization was very interested in an independent viewpoint of the performance of these six managers to use in making promotion decisions.

In the research purpose condition (RESRCH), the raters were told that their performance ratings of these managers were going to be used to help do validation research of a managerial assessment center recently introduced in the sponsoring organization. It was stressed that this assessment center involved a multimillion-dollar investment for the company, and thus, the

ratings were important in providing an independent source of performance ratings for the managers in the work sequences.

In all conditions, the importance of the study as a "$100,000 contract to SUNY-Albany" was emphasized prior to data collection.

In the VT condition, an explanation of the videotape procedure and the rating forms was given. The raters were then shown each of the six videotape sequences, and asked to rate the performance of the managers at the conclusion of each tape as well as completing the confidence ratings. In this condition, all ratings for each videotape were collected before the next tape began so that raters could not change their ratings after seeing several tapes.

In the SRB and NRBS conditions, raters were told that the performance interview between the managers to be rated and the employee were tape-recorded, and then were transcribed into scripts. The raters were told that the employee in the vignettes was actually a member of the Personnel Department who was playing the part of a disgruntled engineering manager. The ratings were to be made on the managers who were providing performance feedback to this employee.

In the SRB condition, raters were told they could refer back to the scripts as often as they wanted while making their ratings. In the NRBS condition, raters were told three times during the initial briefing that they could not refer back to the scripts after they had read each through once. They were instructed to make their ratings after this first reading, and were closely monitored by the experimenter.

In both of the script conditions, the raters had to finish the first script, their performance ratings, and confidence estimates prior to receiving the next script. They had to return their ratings and the script to the experimenter before they received another script. This was done, as with the videotape procedure, to control for the fact that raters might change their ratings after they read or saw several behavioral sequences.

In all conditions, raters completed a questionnaire after finishing their performance ratings. This questionnaire contained items related to the interviewing variables and items designed to assess the raters' understanding of the purpose of measurement and the importance of the study:

a. part of a promotion decision
b. for research validating tests
c. for personal growth and development
d. I don't know

Using a 5-point scale, the importance of the study was assessed with the following question: "To what extent do you feel the performance ratings you completed are important to the sponsoring organization of this study?"

All subjects then received a lecture on how the results of the study in which they had participated were to be used by AFHRL.

Results

Manipulation Checks

The analysis of the single item used to check on the manipulation of the purpose of measurement revealed a significant effect ($p < .0003$) for experimental conditions. Subjects in the administrative (promotion) and research conditions did, in fact, indicate that they were in those conditions. The analysis of the single item used to check on the manipulation of importance of the study revealed no differences across experimental conditions. It was necessary to control for importance as a social force in this experiment since it could pose a threat to internal validity. That is, the results of this study could have been explained by the greater importance of the performance ratings for administrative purposes versus those used only for research. The results of this manipulation check were consistent with the attempt to control for importance as an unintended social force in this study.

Intervening Variables

The a priori clusters of items to measure the four intervening variables were subjected to an internal consistency reliability analysis. The alpha reliabilities, based on 134 respondents, and number of items per scale were: (a) rater motivation - eight items, .77; (b) rater trust - six items, .65; (c) rater acceptance - nine items, .82; and (d) rater confidence - ten items, .83. These reliabilities are acceptable for research under prevailing psychometric standards (Nunnally, 1978). The reliabilities for all three studies of this research project and a study done of rater training that used these scales (Study 2) are contained in Table 1. As can be seen, with the exception of the rater trust variable for studies 3 and 4, all measures of the variables reached acceptable levels of reliability. Rater trust was excluded from analysis in the studies in which it had unacceptable reliability.

Table 1. **Reliabilities for Intervening Variables for Four Studies**

Scale	Items	Study 1	Study 2	Study 3	Study 4
Rater Motivation	8	.77	.80	.76	.72
Rater Acceptance	9	.82	.85	.81	.85
Rater Trust	6	.65	.69	.39	.48
Rater Confidence	10	.83	.80	.83	.80

Note. The sample sizes for the alpha estimates were 134, 88, 111, and 90, respectively, for the four studies. Studies 1, 3, and 4 were part of this research project. Study 2 was reported in more detail in Ruddy (1985).

ANOVA Analyses

Since the hypotheses of this research dealt with relationships between the independent variables and both the intervening and specific dependent variables, separate ANOVAS were computed for the intervening and dependent variables. The results of the ANOVAs for the intervening variables showed no significant main or interaction effects for any of the four

intervening variables. The ANOVA results indicated no significant main or interaction effects for the purpose of measurement.

There were two significant results for experimental technique. As indicated by the mean values for the dependent variables in Table 2, ratings in the VT and NRBS conditions had significantly more ($p < .01$, $w^2 = .06$) range restriction than did those in the SRB condition. For distance accuracy, ratings in the VT condition were more accurate ($p < .05$, $w^2 = .05$) than were ratings in either of the script conditions. There was no difference in distance accuracy between the script conditions.

Table 2. **Means for Significant Findings: Study 1**

Experimental treatment	Dependent variable: Range restriction[a]	Distance accuracy[b]	Leniency[c]
VT	1.45	1.60	
NRBS	1.48	1.72	
SRB	1.60	1.80	
ADMIN			.37
RESRCH			.55

Note. Abbreviations used for experimental treatments are VT = videotape, NRBS = no refer back script, SRB = script refer back, ADMIN = administrative purpose, RESRCH = research purpose.

[a]The higher the mean, the less the range restriction.

[b]The higher the mean, the lower the distance accuracy.

[c]The higher the mean, the more leniency.

Correlational Analyses

The variables relating to the hypotheses in this research were subjected to correlational analysis. This was done to examine the relationship between acquaintance with the job and the dependent variables, and to allow some post hoc analysis of the relationships among all variables. It was also done to examine the effects of the intervening variables on measurement quality. The results for Study 1 are contained in Table 3. As can be seen by examining Table 3, there is little relationship between the "acquaintance with the job" variables from the biographical questionnaire and the quality of the measurement. Of the 30 correlations between the acquaintance with the job variables and the quality of measurement variables, only four were significant. There were no significant relationships between the two accuracy dependent variables and the "acquaintance with the job" variables; however, two of the five relationships between the job acquaintance variables and halo were statistically significant. Ratings by subjects with more performance appraisal and performance feedback experience showed a greater halo effect.

Table 3 also displays the relationships between the job acquaintance variables and the intervening variables. Again, with the exception of the performance feedback experience, the other job acquaintance variables showed a low correlation with the intervening variables. The feedback experience variable demonstrated a significant positive correlation with three of the intervening variables: rater motivation, acceptance, and confidence. The other significant

relationship indicates that amount of supervision experience with engineers is negatively related to rater acceptance.

Table 3. **Correlation Results for Study 1**

	Acquaintance with job			Quality of measurement							Intervening variables			
	SE	PA	FB	SU	DA	LN	HO	RG	CA	CN	RM	RT	RA	RC
WE	30	30	18	70	11	05	10	-01	-12	03	12	05	-08	-02
SE		33	13	46	10	-09	07	02	-07	-10	00	00	-19	-01
PA			52	46	-07	-08	18	-16	01	13	07	-05	-07	08
FB				26	-08	02	22	-10	08	14	26	00	19	21
SU					00	04	10	-07	-02	08	00	-03	-13	01
DA						32	-28	37	-78	10	-02	-09	-05	-10
LN							04	-19	-15	-26	-12	11	-08	-23
HO								-64	09	01	-05	08	-06	04
RG									01	25	20	-03	23	14
CA										-03	10	04	10	13
CN											34	-06	30	61
RM												22	62	54
RT													37	15
RA														62

Note. Decimals are omitted. For n = 134, correlations of .14 and .21 are significant at the .05 and .01 levels, respectively. WE = Work experience, SE = Supervisory experience with engineers, PA = Performance appraisal experience, FB = Feedback experience, SU = Supervisory experience in general, DA = Distance accuracy, LN = Leniency, HO = Halo, RG = Range restriction, CA = Correlational accuracy, CN = Confidence in ratings, RM = Rater motivation, RT = Rater trust, RA = Rater acceptance, RC = Rater confidence.

Finally, in Table 3, the relationships between the intervening variables and the dependent variables show interesting trends. Three of the four correlations between the intervening variables and the confidence variable were significant, indicating the higher the rater motivation, acceptance, and confidence as measured by the post-experimental questionnaire, the higher the confidence the raters reported in their ratings of the performance sequences. However, it should be noted that these are correlations between self-report measures of the same process.

The other significant relationships indicate that the higher the rater motivation, acceptance, and confidence, the more range restriction in the ratings. There was a significant negative relationship between rater confidence and leniency, and a trend for both rater motivation and acceptance to also be negatively related to leniency. This means the more confident the rater, the less lenient the ratings. Finally, there was a trend for rater motivation, acceptance, and confidence to be positively related to correlational accuracy.

Discussion

The results of this study provide some answers to the issues raised in the introduction and raise interesting questions for both current research in performance appraisal and the JPMS project. In terms of the purpose of measurement, there was no support for any of the hypothesized causal relationships in Figures 3, 3a, or 3b. The lack of a main effect of the measurement purpose on the quality of measures dependent variables is consistent with the findings of McIntyre et al. (1984) and inconsistent with the Zedeck and Cascio (1982) results.

It should be noted that this study used the same formulas for the calculation of the accuracy indices as did McIntyre et al. (1984), and as they noted, their measurement of these variables differed from that of Zedeck and Cascio (1982). Another difference between these two previous studies was that one used "paper people" vignettes (Zedeck and Cascio, 1982), while the other used videotapes (McIntyre et al., 1984). The lack of a significant interaction in the current study between the presentation mode of the stimulus material (VT, NRBS, and SRB) and the purpose of measurement partially argues against this interpretation of the different findings in the two previous studies.

It should be noted, however, that the stimulus materials used in this study for the "paper people" vignettes were much different in informational content than those used by Zedeck and Cascio (1982). Their materials were short paragraphs describing the performance of 33 different supermarket checkers, whereas our vignettes were the full scripts used to develop the Borman et al. (1976) videotapes. Thus, it may be that the effects of purpose found by Zedeck and Cascio (1982) are due to the low amount of performance information on the ratees, relative to that in the videotapes, provided by their vignettes. When we equated the informational content between videotapes and vignettes in this study, the manipulation of purpose may not have been strong enough to affect rating accuracy. Obviously, this could also account for the different findings for the previous two studies (McIntyre et al., 1984; Zedeck & Cascio, 1982). Future research needs to better define and address this hypothesized explanation.

Another possible explanation for these findings is that purpose of measurement manipulations in "created" laboratory settings are not effective enough to impact on rating accuracy. There are several points that appear to support this interpretation. Neither this study nor McIntyre et al. (1984) found a significant main effect for leniency, whereas in other studies in "real" situations cited earlier, the administrative purpose condition almost uniformly produced more lenient ratings. The Zedeck and Cascio (1982) study found less discrimination for ratings in the administrative condition; however, they had no measure of leniency. Discriminability and leniency are not the same thing.

Further support for the notion that it may not be possible to manipulate purpose of measurement in contrived situations (i.e., paper people vignettes or videotapes) comes from the lack of any main or interaction effects of purpose of measurement on the intervening variables. It has long been assumed that the reason raters in "real" situations are more lenient when the purpose of the performance rating is administrative is because their motivations are different from raters in research or growth conditions.

In terms of the JPMS project, this final interpretation would indicate that, in field research to validate the ASVAB, the performance ratings should be collected for research purposes. Although this may not affect the accuracy of the ratings, it could impact on the amount of leniency in the data. Obviously, severe leniency, which would cause range restriction in the measurement of job performance, could seriously impact on the ASVAB validation effort.

In terms of the different methods employed to present the stimulus material, the results from this study support the hypothesis that the videotape is superior to either vignette (script) condition. The raters in the VT condition were significantly more accurate (DA) than in either of the script conditions, and showed less range restriction than did raters in SRB conditions.

This finding has important methodological consequences. In research using a created stimulus in the "true score" paradigm to test the impact of either an organizational or individual variable on rating accuracy, the videotape is the more appropriate method. The results of previous research using vignettes must be viewed with caution, and should not be used to make recommendations for changes in performance measurement systems in applied settings. For example, if three different training programs are being evaluated in terms of the one that can best improve rater accuracy, results from a vignette study may not be correct, whereas results from a videotape study may be accepted with more certainty. If significant costs are involved in this decision, it seems rather prudent to use the videotape technique.

Furthermore, this finding has serious implications for both past and future research. One must view with skepticism the results of studies that used the "paper people" approach, until repeated with a videotape technique. Future researchers may want to consider using only the videotape method; however, it must be emphasized that this finding of differences in accuracy between the two methods needs to be replicated before firm advice can be given.

In terms of the JPMS project, the best practical advice would be to use the videotape methodology in future research that examines the characteristics that affect rating accuracy. It should be noted, however, that these different methods will be studied again within this project. The implications of this replication will be discussed later in this paper.

The hypotheses regarding the acquaintance with the job variable, depicted in Figures 4, 4a, and 4b, received little support. There was no support for the hypothesized relationship between job acquaintance and rating accuracy, and raters with more performance appraisal and feedback experience had more halo in their ratings. This latter result may not be surprising if one accepts the argument and empirical evidence that halo is the most common effect found in performance rating data. It would be reasonable, therefore, to assume that persons with more experience in performance appraisal would exhibit more halo in their ratings of job performance. Obviously, this would be an interesting hypothesis to pursue, particularly if one had access to a large data set containing these variables.

The weak support for the hypothesized relationships between acquaintance with the job and the intervening variables also indicates that this independent variable did not have a powerful effect in this study. As noted in the introduction to this report, there is no previous evidence regarding the relationship between acquaintance with the job and rating quality. The evidence that does exist is indirect (Bazerman et al., 1982; Freeberg, 1968; Jackson & Zedeck, 1982; Scott & Hamner, 1975), and never directly addresses the degree of acquaintance the rater has with the actual job the ratee is performing. Given the results of this study, it would appear that acquaintance with the job is less important in affecting rating accuracy than are factors such as acquaintance between the rater and ratee, familiarity with the ratee's previous performance, and degree of responsibility over the rater.

These results must be accepted tentatively, however, because of the nature of the subject sample and the job situation. It may be most raters were familiar enough with the job situation, a performance feedback interview, that additional experience with this job task would not

significantly increment one's rating ability. This would mean that beyond a certain level of familiarity with the job, additional experience would have no effect. It is also true that most of the raters in this study (86%) had some experience with performance appraisal. Therefore, to adequately test this hypothesis, one would have to select a job and subjects such that at least 50% had zero acquaintance with it.

In terms of the JPMS project, these results would indicate that acquaintance with the job may not be a critical factor in terms of measurement quality. It should be noted that this recommendation is being made for raters who have at least some knowledge of the job. It is not necessary to find extremely experienced raters to ensure more accurate ratings; however, a note of caution is necessary. This does not argue that raters with absolutely no acquaintance with the job could provide accurate ratings, as this was not tested in this study.

Finally, the results in Table 3 do provide moderate support for the hypothesized relationships between the intervening variables and the dependent variable as depicted in Figure 2. There is clearly a trend for the intervening variables to be positively related to correlational accuracy, and significantly negatively related to leniency and range restriction. Although one could hardly call this convincing evidence, it does suggest the link between these personal, motivational variables and performance rating quality does merit serious consideration in future research. Furthermore, to our knowledge, this is the first empirical demonstration of this linkage in the literature.

Given this evidence of this linkage in the model, it is unfortunate that the linkages between the independent variables and intervening variables did not appear as hypothesized. If one could establish a firm link between the intervening variables and performance rating accuracy, then research could focus on variables that positively impact on the intervening variables. This would be a more efficient paradigm than one that has to include the measurement quality variables in this study.

In terms of the JPMS project, it seems clear that any change in the system should be examined in terms of its effects on these intervening variables. The questionnaire for these items is quite short and self-administering, and the variables all have acceptable internal consistency reliabilities. These intervening variables will all be included in subsequent research studies in this project.

This study has provided valuable guidelines for the subsequent research in this project. It is apparent, in terms of the use of the Borman (1978) tapes and scripts we have selected, minimal acquaintance with the job is the only qualification needed for subjects. This allows us to broaden our potential subject pool, and reduce the size of our biographical questionnaire. The reduced number of items to measure the intervening variables that resulted from the reliability analyses will also allow us to reduce the length of the questionnaire.

The purpose of measurement findings are at a dead end, and no further research is necessary. It is highly recommended, based on the results of this and earlier studies, that all field studies that collect performance measurement data do so "for research purposes."

Finally, the "paper people" versus videotape controversy has not yet been completely settled; however, the "no refer back script" condition fared poorly, and was dropped from the next experimental study in this project. This reduced the number of subjects needed for the next study and allowed us to increase the power of the design.

III. STUDY 2

This study was concerned with examining the "true score" paradigm for the evaluation of rating accuracy developed by Borman et al. (1976). By identifying a new set of SMEs, this study had the following three purposes: (a) to develop a new set of "true scores" for the videotapes being used in this project; (b) using the original BARS scales as a starting point, to develop a new set of rating scales with performance standards as anchors for the numerical scales; and (c) to explore the criterion deficiency of the current BARS (Borman et al., 1976) for measuring the performance of a supervisor in a performance appraisal feedback interview. All the materials used in this study, including correspondence to the SMEs, are contained in Appendix D.

Method

Participants

Participants were recruited from a local Personnel Association by means of a letter to the membership and followup phone calls. As can be seen from the correspondence to the participants, we were seeking Human Resources Managers who had at least 3 years of experience in completing performance appraisals and conducting performance appraisal interviews. Of our eight participants, all met these criteria, with the minimum experience being 5 years in a supervisory capacity. There were five males and three females, and they held varying positions in Personnel from Director to Compensation Analyst. All participants were paid $50.00 for their help as SME consultants to the project.

Procedure

After identifying the SMEs, each was sent a letter describing the three tasks they were going to perform and the date of the first meeting. The first meeting was spent developing performance standards for the BARS scales (Borman et al., 1976) used previously. This meeting, and all subsequent meetings,were tape-recorded, and a copy of these tapes is available from the principal investigator. All meetings were led by the principal investigator and attended by the project associate, who helped focus the meetings on the three tasks for this study.

During this first meeting, the primary emphasis was on the development of performance standards; and the SMEs were given a general guide as to what was meant by performance standards (see Appendix D). However, we were also concerned with exploring the criterion deficiency of the BARS during this first meeting, since the development of performance standards implied some improvement to the scales. By tape-recording this meeting, gathering of evidence for the criterion deficiency of the BARS, as well as the development of a new rating form with performance standards, was greatly aided. The new rating form, which was used in Study 4, is included in Appendix I.

We had intended to give each SME a copy of the videotapes to review during the interval between meetings; however, the quality of the copies was so poor that this was impossible. We did give the SMEs the BARS, rating forms, and scripts of the videotapes for their review

prior to the next session. We also discussed the rating tasks and the videotapes with them in some detail so that they would have a better frame of reference to review the scripts and rating materials prior to the next meeting. In other words, we were trying, as Borman (1978) did, to maximize their rating performance on the videotapes.

The second meeting of the SMEs began with the ratings of the six manager videotapes used in this project. The SMEs were shown a single sequence of one manager while they had the script and rating forms in front of them. This was the same procedure used by Borman (1978). The SMEs were told we would rerun a videotape if they needed to view it again; however, there were no requests to do so. Again, this meeting was tape-recorded since, as we expected, there were a large number of clarifying questions and considerable discussion about the rating task. This provided a continuing rich source of qualitative data about the criterion deficiency of the BARS scales, the appropriate criterion space for performance in an appraisal interview, and information on performance standards. In fact, the meeting leader used probes to address both performance standards and criterion deficiency issues related to both the videotapes and the BARS content.

After each videotape was completed, each SME made a rating on the performance dimensions described by the BARS, but did not share them publicly. The SMEs were told to study their ratings, and consult the scripts before the next meeting, to arrive at a final rating for each manager on each dimension. They were also told that we would be reaching consensus decisions on the ratings of each dimension for each manager at the next meeting.

At the next meeting, we arrived at consensus for the ratings of each manager on each performance dimension by using a Nominal Group Technique (NGT) (Delbecq, Van de Ven, & Gustafson, 1975). In addition, we collected the ratings each SME had made privately of the performance of the managers on the videotapes. Thus, we had both consensus ratings with zero variance and individual ratings for which we could compute means and variances.

Results

Criterion Deficiency

From the meetings with the SMEs and listening to the tape recordings, it became apparent there were several deficiencies in the BARS as applied to the measurement of effectiveness in a performance appraisal feedback interview. The most glaring of these was that there was no measure of the "maintenance of self-esteem" of the employee in the BARS. The SMEs felt that this should be a separate dimension on which the manager is assessed. However, since we were restricted to the number and names of the dimensions defined by Borman et al. (1976), we attempted to reflect this concern and the other criterion deficiency issues in the new "performance standards" rating scale we developed (see Appendix I) for use in Study 4.

The absence of an opportunity to rate the "maintenance of self-esteem of the employee" as a separate performance dimension meant that the SMEs included an evaluation of this dimension when they rated the managers on the original Borman dimensions. As a result, the SMEs effectively redefined the criterion space of job performance in an appraisal interview. Other performance dimensions that the SMEs felt were missing from the Borman dimensions were: (a) prior planning for the appraisal interview; (b) anticipation and defusing of potential conflict areas; and (c) action planning with Whipker, the subordinate manager, on an ongoing process. In effect, by redefining the criterion space for the performance of their manager, the SMEs

redefined the basis upon which to make their ratings. The impact of this redefinition of the criterion space on the evaluation of the performance of the individual managers on the videotapes was most pronounced during the NGT used to reach consensus. Thus, in redefining the criterion space and its measurement, the SMEs essentially created a measurement situation decidedly different from the original one contained in Borman et al. (1976). The effect of this redefinition on the SME true scores will be discussed below.

Performance Standards

After listening carefully to the tape recordings, several drafts and a final form of the new "performance standards" rating scale format were developed. The modifications to this new form provided greater specificity and attempted to reflect some of the criterion deficiency discussed above. There was also an attempt to establish performance standards in a binary fashion. Each performance level on the scale for each dimension was written in an "all-or-none" manner, in an attempt to provide the rater with a clear choice as to whether the ratee exhibited the behavior specific to a given scalar point. This is similar to a Behavior Observation Scale (BOS) (Latham & Wexley, 1977), in which the rater checks all job performance behaviors that the ratee exhibits on the job. Thus, the rater makes a binary decision, present or absent, in a BOS. However, for comparison with the rating scales used by Borman et al. (1976), we had to create 7-point scales (see Appendix I). In this process of scale development, some of the performance levels lost the binary character we were attempting to achieve with performance standards. It may be that performance standards scales need to exist on an all-or-none 2-point scale as is the case with a BOS. Attempts to create more scalar points may only confuse the raters. More research is needed on this issue.

SME-Derived True Scores

Tables 4, 5, and 6 contain the results of this study on the development of SME-derived true scores, as well as the ones developed by Borman (1978). It is apparent that the true scores developed in this study are significantly different from those developed by Borman, both in terms of level and pattern. Given the results discussed above in terms of the criterion deficiency of the Borman et al. (1976) rating form, and the fact that the SMEs redefined both the criterion space and the measurement of performance in an appraisal feedback interview, this is understandable.

Discussion

Based on the results and observations of this study, it was felt that the SME-derived consensus true scores should be used as our target scores for determining accuracy in ratings; thus, these scores were used for all research in this project. This decision was based on several considerations. First, it has been 10 years since Borman et al. (1976) developed their expert true scores; and the changed true scores may be a result of the time which has elapsed. That is, the definition of what is effective in terms of a performance feedback interview may well have changed over time as a result of changes in the prescriptions contained in the scientific and practitioner literature. Certainly the emphasis our SMEs placed on "maintenance of self-esteem of the employee" is a direct result of the recent emphasis in management training on this aspect of supervisor-subordinate relationships (Sorcher & Goldstein, 1972). In fact, several of our SMEs mentioned specific supervisory training programs that have this emphasis.

Table 4. Intended Performance True Scores (Borman et al., 1976)

Performance dimensions	Managers					
	1	2	3	4	5	6
Structuring the Interview	5.0	2.5	6.0	4.5	6.0	2.5
Establishing Rapport	2.5	5.5	4.5	5.0	4.0	1.0
Reacting to Stress	1.5	4.5	5.0	4.0	6.5	4.0
Obtaining Information	3.5	3.5	6.0	6.5	3.5	5.0
Resolving Conflict	1.5	2.0	6.0	4.5	4.5	3.0
Developing the Employee	2.5	3.5	3.5	7.0	4.0	2.0
Motivating the Employee	2.0	5.0	5.0	5.5	3.5	2.5

Table 5. Actual Performance True Scores (Borman et al., 1976)

Performance dimensions	Managers					
	1	2	3	4	5	6
Structuring the Interview	2.79	2.79	6.92	4.54	4.38	3.08
Establishing Rapport	1.50	5.93	3.62	5.23	3.08	1.38
Reacting to Stress	3.57	5.00	5.38	4.92	5.15	1.85
Obtaining Information	2.36	4.21	6.15	5.69	2.69	1.54
Resolving Conflict	2.07	4.07	5.62	4.31	2.85	2.08
Developing the Employee	2.71	3.07	3.38	6.62	4.54	1.38
Motivating the Employee	2.29	4.86	4.62	6.15	2.77	2.08

Table 6. Subject-Matter Expert Performance True Scores

Performance dimensions	Managers 1	2	3	4	5	6
Structuring the Interview	6.0	6.0	6.75	2.0	2.0	4.0
Establishing Rapport	4.0	6.5	4.0	4.0	3.0	3.0
Reacting to Stress	5.0	5.5	4.0	3.0	4.0	5.0
Obtaining Information	3.5	5.5	4.0	3.0	2.5	3.5
Resolving Conflict	5.0	5.0	5.0	2.75	3.0	4.0
Developing the Employee	6.5	7.0	4.5	3.0	2.0	5.0
Motivating the Employee	5.0	6.5	5.5	4.0	3.5	3.5

Another major consideration was that our SMEs were really more "expert" than Borman's "experts." Borman et al. (1976) used primarily Industrial Psychologists as his experts, not practitioners of performance appraisal. We feel that practicing Personnel Managers with specific expertise in performance appraisal feedback interviews are simply better judges of the effectiveness of the actors in the Borman tapes than are academic Industrial Psychologists who only write about performance appraisal feedback interviews. Finally, the Borman et al. (1976) true scores represent mean score with a range, thus making it difficult to determine what the best "point" estimation of the population true score is; whereas we have avoided that problem with the derivation of the consensus true scores via the NGT (Delbecq et al., 1975).

These considerations, however, raise the issue of the generalizability of "expert" true scores for use in rating accuracy research. Who are the "true" experts to define effectiveness of job performance in any situation? Is it the managers of the firm? The Personnel Department? Or, technical experts like Industrial Psychologists? Or, does it really matter?

Future research must address this issue of the "trueness of true scores" before we can proceed in further scientific research on the causes of accuracy in performance ratings.

One solution in the empirical literature to this problem has been to use the "mean" scores of the performance ratings of the subjects in the accuracy research study as the "true" scores for the derivation of the accuracy indices. This may well solve the question of the validity of the true scores for a given subject pool, but it creates a "monster" in terms of generalizability. Each set of true scores in this procedure is unique to the subject pool in which they were developed. Using them to compute accuracy scores in another study with a different subject pool is totally erroneous without first determining if the distributions of true scores for the two sets of subjects are the same. This has not been done in the literature that uses mean scores

of the subjects to define the "true" score matrix for the computation of accuracy indices. One might legitimately ask if this type of scores can truly be called "true" scores.

This methodological issue regarding the "trueness" of true scores raises rather serious questions about the results of the numerous research studies in this field over the past decade. If there is more than one set of true scores for either a videotape or vignette, be it for managers, recruiters, or college lecturers, will the results and subsequent interpretations and recommendations for action differ as a function of the set of true scores used? For example, are the memory effects on rating accuracy recently uncovered by Murphy and Balzer (1986) a function of the specific set of true scores generated by their experts, 13 graduate students? If a different set of scores were generated by other experts, would the results be the same? This issue becomes more crucial when organizational interventions and changes are based on the results of rating accuracy research; e.g., in the recommendation of one rater training program over another.

Not only is this a serious issue for future research on rating accuracy, but this line of reasoning has important implications for any rating accuracy research done within the JPMS project. Which is the correct set of true scores upon which to compute the accuracy indices? Who are the appropriate experts to derive the true scores? In this research project, we are satisfied, at this point in time, that we have used an appropriate group of SMEs to derive the true scores for the Borman et al. (1976) videotapes. If new videotapes are to be used in rating accuracy research, researchers must be certain to establish true scores based on SMEs using the procedure described for this study, particularly in the development of a performance measurement system for enlisted specialties.

In terms of the other results of this study, it is apparent that there are some problems with the use of the Borman tapes. The scripts are good, but the actors are out-of-date in terms of dress and slang expressions. Further, the results of this study indicate that the original rating scales (Borman et al., 1976) need to be modified in light of the criterion deficiency issues. In terms of the "maintenance of the self-esteem of the employee" dimension, it will be necessary to rewrite the scripts to reflect this performance dimension more sharply. Finally, this criterion deficiency issue may have shown the weakness of any single measurement method, ratings from a single source, to completely measure the job performance of an individual. The performance ratings of the managers by our SMEs reflect one perspective on the measurement of performance of the managers. This "criterion deficiency" problem must be attended to in the development of a performance measurement system to validate the ASVAB.

IV. STUDY 3

Study 3 in this research project was an extension and partial replication of Study 1. We extended Study 1 by dropping the "administrative" purpose condition, and used only the "research only" condition as will be done in the JPMS project. This study also focused on a replication of the results related to the stimulus material (videotape versus vignette) and an examination of the hypotheses related to quality of instructions on the rating scale in terms of the level of detail as contained in Figures 6, 6a, 6b, and 6c. Since we had discovered that the acquaintance with job variable was not related to rating accuracy on the videotapes of managers conducting performance appraisal interviews, it was not necessary to control for this variable through subject selection. However, it was measured in this study to continue to test the hypothesized relationships in Figure 2.

Method

Experimental Design

A completely randomized, 3 X 2 factorial, fixed effects design was used to collect the data to test the hypotheses. This allowed for three levels of the first factor, level of detail, and two levels of the second factor, type of stimulus material (or, experimental paradigm).

For the first factor, level of detail, three levels were used in an attempt to reflect the range of instructions that should accompany a rating form. We emphasize the word "should" since we did not include a "strawman" or placebo condition (i.e., one with such terrible instructions on the rating form that it would be difficult even to figure out where the ratings belonged). The lowest level of detail (LOLEV) included an introduction to the experimental task in terms of rating the videotapes, with a brief description of how to use the rating scales. The moderate level of detail (MODLEV) was the standard set of instructions Borman et al. (1976) developed for use with the videotapes. This included guidelines for making performance ratings and some cautions against common rating "errors." The high level (HILEV) version included both a more detailed, step-by-step set of instructions and some modifications to clarify the cautions contained in the MODLEV instructions. These three sets of instructions are contained in Appendices E, F, and G.

The second factor, experimental method, had the two conditions from Study 1 that were being replicated in this design. This included the videotape (VT) and the script refer back (SRB) conditions which have already been described.

Subjects

Data were collected from 111 students who were recruited from the general student population at SUNYA and received payment of $10.00 for their participation. Although only 90 subjects were necessary for sufficient power, given the experimental design (Cohen & Cohen, 1975), we did not have complete control over the number of subjects who would show (or not) for an experimental session even though we used a sign-up sheet. The data from all subjects were used to estimate the reliabilities of the intervening variables. For the analyses that tested the hypothesized main effects and interactions, it was necessary to randomly eliminate subjects from some cells to achieve equal cell numbers so that the expected mean squares could be correctly estimated. This resulted in 16 subjects per cell.

Research Variables

There were two independent variables in this study. The first independent variable, level of detail, was manipulated by creating three sets of instructions that varied in length and clarity. These were created by the members of the research team through extensive discussions and re-drafting of the three sets of instructions until all team members were satisfied that differing levels of detail were represented. It was decided not to use outside judges in this task because it was not clear who the "experts" would be for this task. Further, the concept of detail of instructions for a rating scale of job performance was a complex one that involved a knowledge of the videotapes, the rating tasks, and the BARS rating scale used (Borman, 1978) in this

method. It was felt the research team was probably as good a set of experts as could be found.

The second independent variable, the experimental technique, was the same two conditions (VT and SRB) used in Study 1. The no refer back script condition was dropped because it fared the poorest in terms of the results of Study 1. Data were also collected on the acquaintance with the job variable used in Study 1 to cross-validate those results.

The intervening variables were the same as used in Study 1. The dependent variables used in Study 1 were also assessed in this study. In addition, using the "SME-derived" target scores described in the section on Study 2, four dependent variables were created using the same formulas from McIntyre et al. (1984) discussed in Study 1 of this paper. Thus, we were able to analyze this study using both the Borman (1978) true scores and those derived from our SMEs in Study 2.

In addition, we created dependent variable measures of the effectiveness of the level of detail in the scale instructions. Items were written to measure the subjects' evaluations of the "quality of the form used" (Quality) and how well the rating form helped them to understand the rating task (Understand). These additional items are contained in the post-experimental questionnaire in Appendix H, and are marked "Quality" or "Understand."

Although these measures might be seen as tests of the creation of the experimental levels of the detail of instructions variable, we also felt they measured important practical considerations--highly relevant to the JPMS project--of the administration of any set of job performance ratings. Thus, we treated them as dependent variables, but not directly as part of the testing of the hypothesized model in Figure 2.

Experimental Procedure

The subjects signed up for one of the six experimental conditions without any knowledge of the condition, thus assuring a random allocation. After arriving for the experimental session, the students were initially briefed on the general purpose of the research, the importance of the data collection, and their role. No experimental conditions were introduced at this time, except to introduce the importance and purpose of the ratings variables. All subjects were told that the study was a "$100,000 project awarded to SUNY-Albany to rate the performance of managers in a performance appraisal interview situation." They were also told that their ratings "were being used in this research project to examine the validity of a set of tests used in an assessment center by the sponsoring organization." The subjects were given a brief, non-technical explanation of what it meant to validate tests of an assessment center, with strong emphasis on the use of their ratings "for research purposes only." Given that they would be asked to rate the performance interview skills of five managers on the videotapes, this seemed to be a plausible explanation for the purpose (research only) manipulation. After this brief introduction, subjects were asked to sign an "informed consent form," which they all did.

After completing the consent forms, the subjects were asked to complete a brief biographical information form (contained in Appendix H) that was used to assess the acquaintance with the job variable. This questionnaire was a shorter form of the one used in Study 1; however, it contained the same questions we used in Study 1 to assess the acquaintance with the job variable.

After completing this form, the subjects were briefed again on the study, its importance, and the research only purpose. The subjects then received the BARS rating scales (Borman, 1978), a set of forms on which to make their ratings, and a set of instructions on how to use the job performance rating materials with the videotapes. Depending on condition, the subjects received the LOLEV, MODLEV, or HILEV instructions, and were placed on either the VT or SRB condition.

In all conditions, subjects were invited to ask questions to help clarify their tasks. In all conditions, there were some questions about the procedure. After all questions were answered, the research importance of the study as a "$100,000 contract to SUNY-Albany for research only" was emphasized prior to data collection.

In the VT conditions, an explanation of the videotape procedure and the rating forms was given. The subjects were then shown each of the six videotape sequences and asked to rate the performance of the managers at the conclusion of each tape, as well as to complete the confidence ratings. In this condition, all ratings for each videotape were collected before the next tape began, in order that subjects could not change their ratings after seeing several tapes.

In the SRB condition, subjects were told that the performance interviews between the managers to be rated and the employee were tape-recorded, and then were transcribed into scripts. The subjects were told that the employee receiving the appraisal interview was a member of the Personnel Department who was playing the part of a disgruntled engineering manager. The ratings were to be made on the manager who was providing the performance feedback to this subordinate manager. Subjects were told they could refer back to the scripts as often as they wanted while making their ratings. The subjects had to finish the first script, their performance ratings, and confidence estimates prior to receiving the next script. They had to return their ratings and the script to the experimenter before they received another script. This was done, as with the videotape procedure, to control for the fact that subjects might change their ratings after they read several behavioral sequences.

In all conditions, subjects completed a questionnaire after finishing their performance ratings. This questionnaire contained items pertaining to the intervening variables, and to the quality of rating form dependent variables (Appendix H). All subjects then received a lecture on how the results of the study in which they had participated were to be used by AFHRL.

Results

Intervening Variables

Based on the analyses from Study 1, the questionnaire to measure the intervening variables was reduced. The four scales were subjected to internal consistency reliability analyses. The alpha reliabilities, based on 111 respondents were: (a) rater motivation, .76; (b) rater acceptance, .81; (c) rater trust -.39; and (d) rater confidence, .83. With the exception of rater trust, these reliabilities reached acceptable levels for research (Nunnally, 1978). Since the reliability for rater trust did not reach an acceptable level, it was dropped from further analyses.

MANOVA Results

A 3 (level of detail) by 2 (experimental technique) multivariate analysis of variance (MANOVA) was computed. The Hotellings test was significant $p < .0001$), indicating that there were significant effects for the independent variables.

ANOVA Results

Since the hypotheses of this study dealt with relationships between the independent variables and both the intervening and dependent variables, separate ANOVAs were computed for these two sets of variables. In addition, the two quality of rating form dependent variables (Quality and Understand) were analyzed separately.

Intervening Variables. In terms of the experimental technique, SRB versus VT, there were two significant findings. As seen in Table 7, rater acceptance was significantly higher in the SRB condition ($p < .005$, $w^2 = .07$), and rater confidence was significantly higher ($p < .05$, $w^2 = .03$) in the SRB condition.

Table 7. **Means for Significant Findings for Intervening Variables: Study 3**

	Intervening variable		
Experimental treatment	Rater acceptance	Rater confidence	Rater motivation
VT	30.09	28.83	
SRB	33.63	30.38	
LOLEV			27.76
MODLEV			27.56
HILEV			25.41

Note. Abbreviations used for experimental treatments are VT = videotape, SRB = script refer back, LOLEV = low level of instructions, MODLEV = moderate level of instructions, HILEV = high level of instructions. For all three intervening variables, the higher the mean, the higher the perceptual evaluation.

In terms of level of detail in the instructions, rater motivation was significantly higher ($p < .005$, $w^2 = .07$) in both the LOLEV and MODLEV conditions compared with the HILEV condition (Table 7). There was no significant difference between the LOLEV and MODLEV conditions on rater motivation, and there were no significant interactions for any of these intervening variables.

Dependent Variables. We were able to calculate two scores for leniency, halo, correlational accuracy, and distance accuracy based on the Borman (1978) true scores and our SME-derived true scores from Study 3. We will annotate the results with either (Borman) or (SME) to

indicate which score is being used. Means for the significant findings on the dependent variables are contained in Table 8.

Table 8. **Means for Significant Findings for Dependent Variables: Study 3**

Experimental treatment	Dependent variable: Range restriction[a]	Distance accuracy[b]	Correlational accuracy[c]	Halo[d]	Quality[e]	Understanding
VT	1.50	1.25,1.65	.38	-6.33	17.77	11.39
SRB	1.63	1.44,1.78	.31	-8.74	19.77	12.17
LOLEV			.35			
MODLEV			.42			
HILEV			.29			

Note. Abbreviations used for experimental treatments are VT = videotape, SRB = script refer back, LOLEV = low level of instructions, MODLEV = moderate level of instructions, HILEV = high level of instructions.

[a]The higher the mean, the less the range restriction.

[b]The higher the mean, the lower the distance accuracy. The first mean is for SME-derived scores and the second is for Borman-derived scores.

[c]The higher the mean, the more correlational accuracy.

[d]The lower the mean, the greater the halo.

[e]The higher the mean for both Quality and Understanding of instructions, the higher the perceptual evaluation.

There were significant results for the test of the experimental technique. Ratings in the VT condition were significantly better ($p < .0005$, $w^2 = .12$) in distance accuracy (SME), as seen in Table 8, than were ratings in the SRB condition. Likewise, ratings in the VT condition were significantly better ($p < .01$, $w^2 = .05$) in correlational accuracy (SME) than were ratings in the SRB condition. Ratings in the VT condition were significantly better ($p < .05$, $w^2 = .03$) in distance accuracy (Borman) than were ratings in the SRB condition. Ratings in the VT condition had significantly less ($p < .0005$, $w^2 = .09$) halo (SME and Borman) than did ratings in the SRB condition. However, ratings in the SRB condition had significantly less range restriction than did ratings in the VT condition ($p < .05$, $w^2 = .04$). Finally, on the two added dependent variables, subjects in the SRB condition rated both the Quality ($p < .001$, w2 = .09) and Understanding ($p < .05$, $w^2 = .05$) of the rating form and instructions higher than did subjects in the VT condition.

In terms of the level of detail of instructions variable, ratings in the MODLEV conditions had significantly higher correlational accuracy (SME) ($p < .005$, $w^2 = .10$) than did ratings in the HILEV condition. Ratings in the LOLEV and MODLEV conditions did not differ significantly. Importantly, neither Quality nor Understanding of the rating form was significantly different for the three level of detail conditions.

In addition to these main effects, there were three significant interactions. First, there was a significant interaction ($p < .0005$, $w^2 = .10$) for distance accuracy (SME), as seen in Figure 7. For distance accuracy, lower scores are better. Thus, the interaction is primarily caused by subjects in the HILEV, SRB conditions, whose ratings had the poorest distance accuracy. It is interesting to note that there are only minor differences across the levels of detail in the videotape conditions, and the scores are lower (i.e., accuracy was greater) than for the SRB conditions.

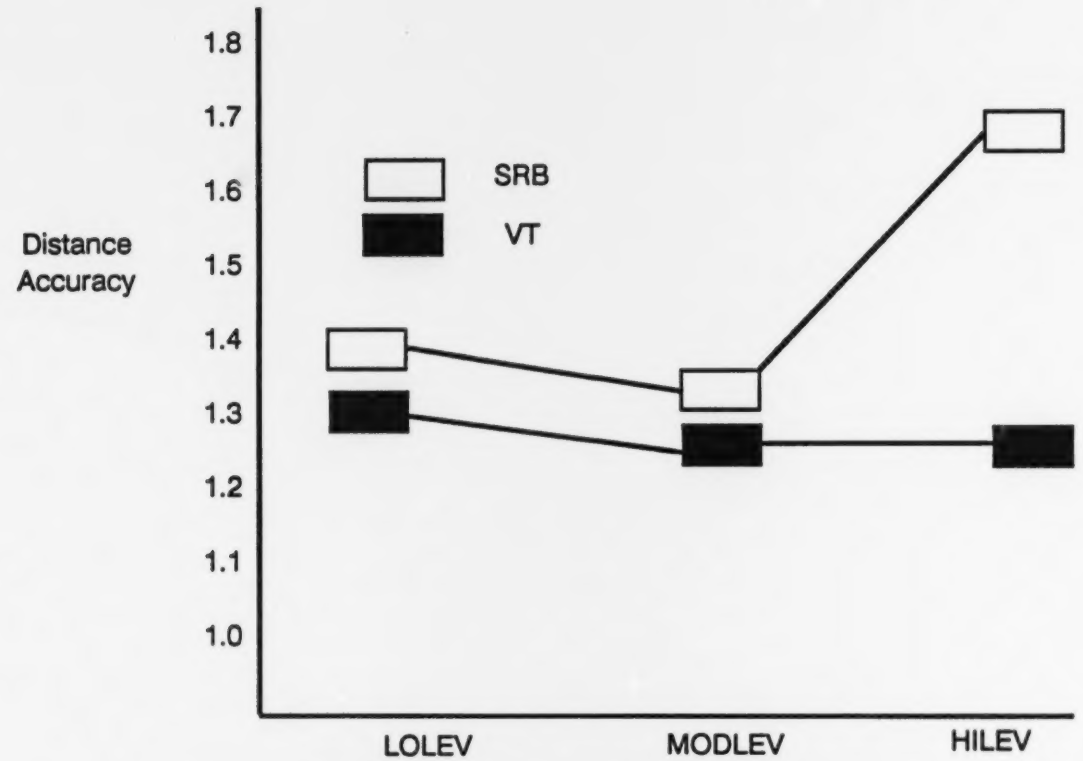

Figure 7. **Interaction for Distance Accuracy (SME): Study 3.**

The results for correlational accuracy (SME) also revealed a significant interaction ($p < .001$, $w^2 = .14$) as seen in Figure 8. With correlational accuracy, the higher the value, the better. Thus, again, it is the HILEV, SRB condition that leads to the interaction because of its low correlational accuracy. Again, there are only minor differences across the level of detail conditions with the videotape.

The third significant interaction was for correlational accuracy (Borman) ($p < .005$, $w^2 = .08$) as depicted in Figure 9. The interaction here is due to the low level of correlational accuracy in the MODLEV, VT condition, and the linear relationship between level of detail and correlational accuracy in the SRB conditions.

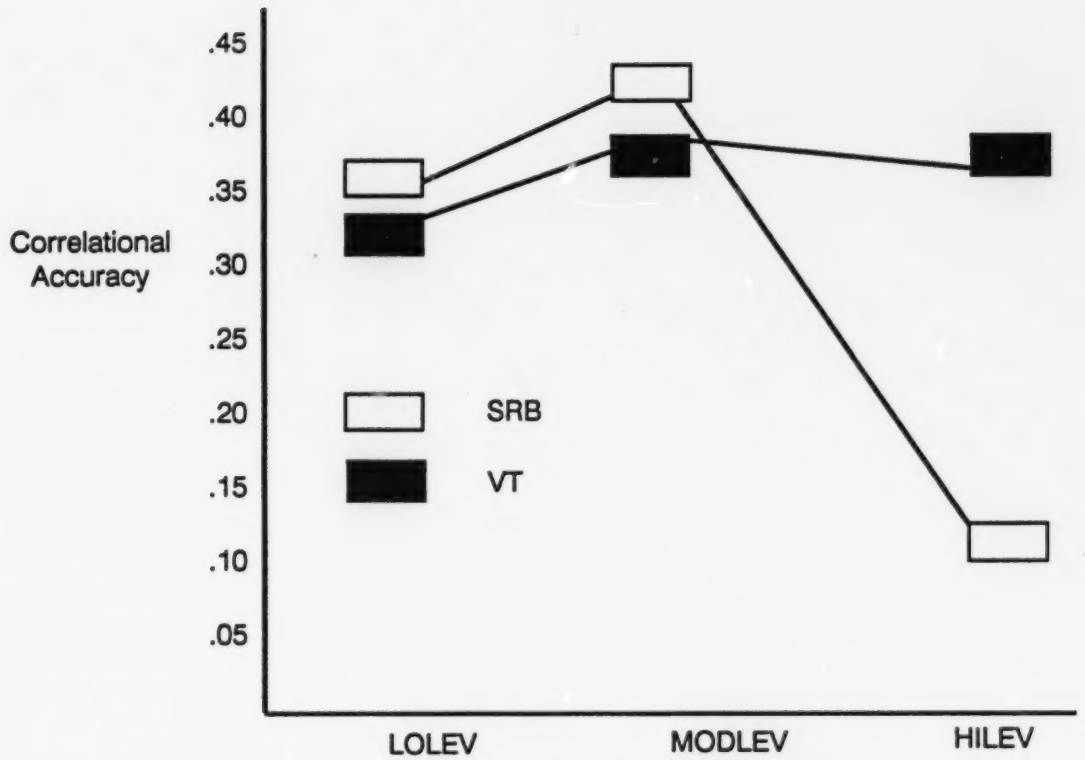

Figure 8. Interaction for Correlational Accuracy (SME): Study 3.

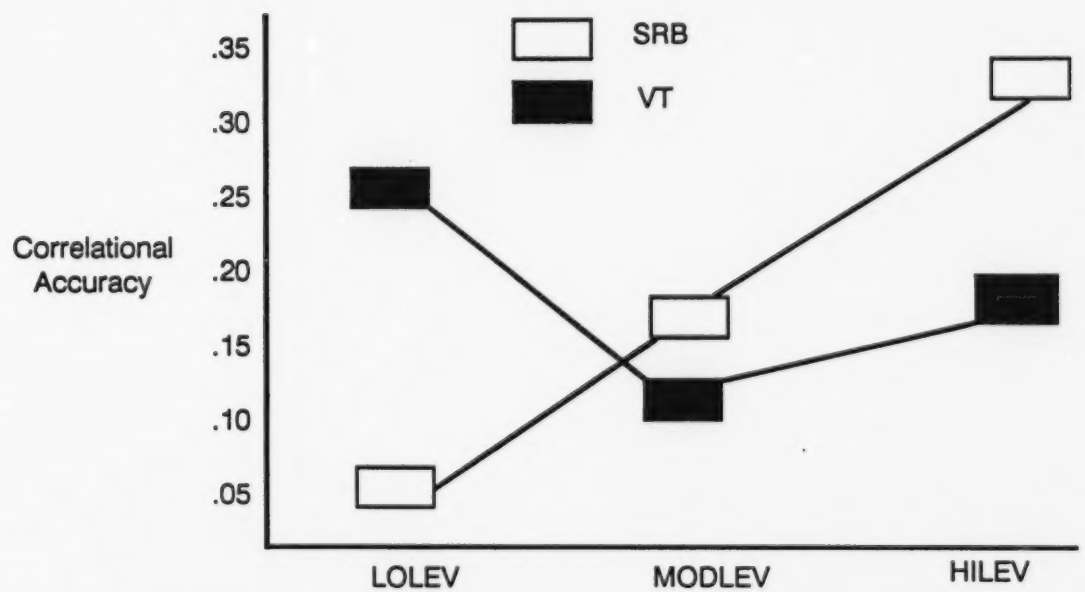

Figure 9. Interaction for Correlational Accuracy (Borman): Study 3.

These three interactions reveal that subjects in the SRB, HILEV condition performed the poorest in accuracy when the SME-derived scores are used, and did best when the Borman scores are used. However, it is important to note that these three interactions are quite consistent for the videotape conditions. Although there is some *fluctuation by level of detail for correlational* accuracy (Borman), there are small differences across levels of detail when the videotape technique is used.

Correlational Results

The correlational results are presented in Table 9. In terms of the acquaintance with the job variables (WE, PA, FB, SU), there is a consistent, negative relationship with correlational accuracy (Borman) and a positive relationship with confidence in the ratings. The latter finding is consistent with the finding from Study 1, whereas we have no explanation for the former finding. It is interesting to note that although there are significant relationships with the Borman-derived accuracy measures, there are no significant relationships with the SME-derived measures. Of the 12 relationships between the acquaintance with the job variables and the intervening variables, only one is significant.

Table 9. **Correlation Results for Study 3**

	Acquaintance with job			Quality of measurement								Intervening variables		
	SU	PA	FB	DA1	CA1	LN	HO	RG	CN	DA2	CA2	RM	RA	RC
WE	62	26	25	16	-24	03	-03	-01	03	00	-04	05	-03	-14
SU		61	67	17	-23	00	-01	01	17	01	-04	09	-12	05
PA			77	04	-13	04	14	-12	20	-05	-01	01	19	04
FB				08	-16	02	08	-05	24	-05	00	-06	-08	17
DA1					-81	22	-37	31	14	29	13	26	13	14
CA1						01	14	01	-01	12	-22	-08	04	03
LN							04	-13	05	24	-16	12	07	03
HO								-73	-03	-41	09	-17	-01	-01
RG									04	51	-06	-14	08	10
CN										12	08	46	44	71
DA2											-64	11	16	25
CA2												23	07	04
RM													55	45
RA														52

Note. Decimals are omitted. For n = 111, correlations of .16 and .22 are significant at the .05 and .01 levels, respectively. WE = Work experience, PA = Performance appraisal experience, FB = Feedback experience, SU = Supervisory experience in general, DA1 = Distance accuracy (Borman), CA1 = Correlational accuracy (Borman), LN = Leniency, HO = Halo, RG = Range restriction, CN = Confidence in ratings, DA2 = Distance accuracy (SME), CA2 = Correlational accuracy (SME), RM = Rater motivation, RA = Rater acceptance, RC = Rater confidence.

The relationships between the intervening and dependent variables show some interesting findings. All three intervening variables (rater trust was dropped due to low reliability) show a strong relationship with the confidence the subjects had in their job performance ratings of the videotapes and scripts. This is consistent with Study 1, and again, it is a self-report, self-report relationship. Of the six correlations between the three intervening variables and the two distance accuracy measures, all are positive, and three are significant. There appears to be a positive relationship here that supports the general hypotheses of this study. Rater motivation and correlational accuracy (SME) are significantly related in the direction hypothesized, and halo and rater motivation are significantly negatively related as hypothesized.

Discussion

The results of this study, along with those of Study 1, strongly support the use of the videotape technique in rating accuracy research, and cast even stronger suspicion on previous research findings that have used the "paper people" vignette technique. The ratings in the videotape condition were significantly more accurate in terms of both distance accuracy (SME and Borman) and correlational accuracy (SME), which, in our opinion, are the most critical dependent variables in rating accuracy research. Further, ratings in the videotape condition showed significantly less halo effect (SME and Borman), a fact that should further the case for this technique.

We are not ignoring the fact that the subjects in the script (paper people) condition responded that they had greater confidence and acceptance, as well as judging the quality and understanding of the rating process higher, than did subjects in the videotape condition. Further, subjects in the SRB, HILEV condition had high correlational accuracy (Borman) as seen in Figure 9. However, these findings are likely due to the greater familiarity that college students would have with a judgment (rating) task involving written rather than videotape stimulus materials. Reading comprehension tests, for example, require similar judgmental processes to the SRB condition in this study. We feel this familiarity explanation would appear to account for the perceptions of the subjects that the conditions with written materials were easier to understand and more motivating. In the SRB, HILEV condition, the subjects were given the structure, through the detailed, step-by-step instructions, to improve their scores over the less structured LOLEV and MODLEV instructions. The true test, however, is that the videotape conditions appeared to have conveyed more information given the other strong accuracy results.

These results underscore and amplify the recommendations made for the JPMS project on the basis of Study 1. The videotape technique is the only acceptable method for examining important personal, organizational, or system characteristics to be included in the performance measurement system that is to be used to validate the ASVAB. Use of the "paper people" technique could easily lead to erroneous conclusions regarding important design features of the measurement system, a situation clearly to be avoided.

The hypotheses contained in Figures 6, 6a, 6b, and 6c received some support from this study, and the results provide some guidance for the JPMS project. It is not clear, however, which level of detail is always best for instructions. Even though the HILEV of detail led to significantly lower correlational accuracy (SME) compared to LOLEV and MODLEV, it is apparent from Figure 8 that this was due to subjects in the SRB rather than the VT conditions. From Figures 7 and 8, subjects in the VT conditions did equally well on correlational accuracy and distance accuracy (SME), and it is clear the main effect for level of detail was due to the poor performance of subjects in the SRB condition with HILEV instructions.

However, complicating these results, as shown in Figure 9, is the fact that subjects in the SRB, HILEV condition achieved the best correlational accuracy (Borman), although not significantly better than subjects in the LOLEV and HILEV videotape conditions. As will be discussed in Study 3, we have much less confidence in the Borman "true scores" than the SME-derived ones in terms of the current, expert opinions on the job performance of the managers in the videotaped appraisal interviews. It would appear that the level of detail of the instructions that accompany the rating form will affect rating accuracy in the "paper people" technique, but have little effect in the videotape method. This finding has serious implications for research that uses the former approach. Not only have we shown in this research project that the videotape technique is superior to the vignette one, but it appears that the results of research using the vignette technique could be further complicated by the instructions that accompany the form. In terms of college student subjects, which much of the previous research has used, the "familiarity hypothesis" seems more plausible, and deserves careful research in the future. As with reading comprehension tests, it may be that more detailed and clearer instructions can improve the performance of college students in rating tasks using vignettes.

Given the demonstrated superiority of the videotape technique in this research, the level of detail of the instructions that accompany the rating form may be irrelevant. There are no significant differences in level of detail for the VT conditions (Figures 7 and 8). It may be that the addition of "step-by-step instructions," "guidelines," and "things to guard against" instructions (see MODLEV and HILEV in Appendices F and G) simply does not improve rating accuracy, and that the simple, straightforward approach in LOLEV is all we need. The HILEV and MODLEV instructions do not harm accuracy compared to the LOLEV instructions; however, they add paper to the rating form. This could be a serious cost consideration in the massive data collection effort that will be necessary to validate the ASVAB.

It is interesting that the HILEV and MODLEV instructions reflect what we would describe as "good practice based on research" for the design of a performance appraisal rating form. Our education as Industrial/Organizational Psychologists emphasized that we should avoid the "traits only, graphic rating" scales that have been clearly shown to be inferior to other, more behaviorally anchored scales. It may be this perceived need for greater specificity has led to the HILEV and MODLEV types of instructions. It could be that with a well-developed, behaviorally anchored rating scale, only very simple instructions are necessary to complete the form. This would certainly be consistent with earlier arguments for the content of rating scales (Kavanagh, 1971).

As a caveat, we would urge, however, that the nonsignificance of results for level of detail of instructions for the videotape conditions not be over-interpreted. Although the additional "guidelines" contained in both the HILEV and MODLEV instructions did not directly impact on rating accuracy, they may serve an arousal purpose. With college students in an already high demand situation created by the experimental setting, this arousal may not have any effect. However, with real raters, as will be used in the JPMS project and the ASVAB validation, this arousal may be necessary. We are, on the basis of these results, unwilling to conclude that very low level instructions would be effective in eliciting accurate job performance ratings in field research. The additional verbiage with the HILEV and MODLEV instructions may be necessary to "set the stage" for raters who have interrupted their daily work to complete performance ratings. This means the findings of this experimental research on level of detail will probably not generalize to field settings. Further research in field settings is necessary to test the effects of level of detail of instructions on measurement quality.

The correlational results replicate some of the results for Study 1, and thus provide stronger support for the model in Figure 2. The relationships between the intervening variables and the accuracy measures, along with those from Study 1, indicate this linkage does exist. Although significant, the relatively small effects represented by these relationships may indicate either that there is some "noise" in the conceptual model (for example, the existence of a third variable impacting on this relationship), or the relationship is simply not as strong as hypothesized. If the latter explanation is true, we may have to question the practical implications of these results. Since these relationships were to be tested again in Study 4, we decided to defer a decision until then.

In terms of the JPMS, several conclusions seem warranted. First, when doing rating accuracy research, only the videotape technique should be used to evaluate characteristics of the performance measurement system under development. Second, the level of detail of the instructions with the rating form does not appear to seriously impact accuracy in the videotape condition, and thus, the most cost-effective approach should be used in the JPMS project. However, we repeat and emphasize our caution that it may be necessary to "set the stage" for the raters when collecting data in the field. The simple instructions used in the LOLEV condition in this high demand experimental setting may not work in the field. Third, the intervening variables appear to be important in terms of their influence on rating accuracy, and thus, the development of the performance rating system should be concerned with the impact of alternate designs on the variables of rater motivation, acceptance, and confidence. The scales we have developed to measure these variables should be used in the continuing research efforts within the JPMS project to evaluate optional features of a performance measurement system.

V. STUDY 4

Study 4 in this research project was an extension of the earlier work in Studies 1, 2 and 3, but particularly concerned with testing the hypotheses contained in Figures 5, 5a, and 5b with regard to a rating scale with performance standards versus one with a BARS format (Smith & Kendall, 1963). Based on the results of the earlier studies, we used the "research only" purpose condition (Study 1), the videotape technique (Studies 1 and 3), and the HILEV instructions (Study 3); and we scored the accuracy variables using both the Borman and SME-derived true scores (Studies 2 and 3). This study was also concerned with exploratory research on the mode of data collection for performance ratings. All of the materials used in this study are contained in Appendices G, H, I, and J.

Method

Experimental Design

A completely randomized, 3 X 2 factorial, fixed effects design was used to collect the data to test the hypotheses. This allowed for three levels of the first factor, mode of data collection, and two levels of the second factor, rating scale format.

The first factor, mode of data collection, consisted of three experimental conditions. The first one was "experimenter present, verbal instructions (EPVI)." In this condition, the experimenter explained the procedures for the rating task, discussed the set of rating instructions (HILEV from Study 2), and offered to answer any questions the subjects had regarding the rating

procedures. In the second condition, "experimenter present, written instructions (EPWI)," the experimenter distributed a set of written instructions (HILEV) but gave no verbal instructions, and naturally, answered no questions. In the third condition, "videotape experimenter, verbal instructions (VEVI)," the experimenter appeared on videotape to provide a verbal explanation of the rating task, and as in the EPVI condition, covered the rating instructions (HILEV). No questions were allowed in this condition.

These three conditions were chosen since they represent three ways performance rating data can be collected in the JPMS project, and later in the validation of the ASVAB. It is important to determine the most accurate and cost-effective manner to collect the performance appraisal data to validate the ASVAB. If accuracy were equal across these conditions, the most cost-effective mode would be to simply include a good set of instructions for the completion of the rating form, without any elaborate data collection procedures such as training or the use of experts to assist in the completion of the rating forms.

The second factor, rating scale format, consisted of the use of the BARS format developed by Borman et al. (1976) versus the use of the performance standards format developed by our SMEs in Study 2.

Subjects

Data were collected from 90 students who were recruited from the general student population at SUNYA, and who received payment of $10.00 for their participation. This provided us with sufficient power for the experimental design (Cohen & Cohen, 1975).

Research Variables

There were two independent variables in this study. The first independent variable, mode of data collection, was manipulated by using the HILEV instructions from Study 3 in combination with three different ways of collecting the performance rating data. These three modes will be explained further in the procedures section. Since there were no differences in rating accuracy due to different levels of detail in the instructions in the videotape conditions in Study 3, we felt free to choose any of the three levels for this study. It was felt by the research team that the HILEV instructions contained all the information of the MODLEV but were somewhat clearer. The LOLEV was eliminated because of the written instruction only condition (EPWI). In Study 3, subjects were permitted to ask questions; however, in this study, this was not permitted in the EPWI condition. It was felt that the LOLEV instructions were inappropriate for a written only situation.

The second independent variable, rating scale format, was created by using the BARS format from Borman et al. (1976) versus the performance standards format we created in Study 2.

Data were also collected on the acquaintance with the job variable used in Studies 1 and 3, to attempt to clarify its relationship to the other variables in this research project.

The intervening and dependent variables were the same as used in Studies 1 and 3. In addition, the newly created dependent variables (Study 3) concerned with the quality and understanding of the rating scale were also measured.

Experimental Procedure

The subjects signed up for one of the six experimental conditions without any knowledge of the condition, thus assuring a random allocation. After arriving for the experimental session, the students were initially briefed on the general purpose of the research, the importance of the data collection, and their role. No experimental conditions were introduced at this time except the importance and purpose of the study. All subjects were told that the study was a "$100,000 project awarded to SUNY-Albany to rate the performance of managers in a performance appraisal interview situation." They were also told that their ratings "were being used in this research project to examine the validity of a set of tests used in an assessment center by the sponsoring organization." The subjects were given a brief, non-technical explanation of what it meant to validate tests of an assessment center, with strong emphasis on the use of their ratings "for research purposes only." Given that they would be asked to rate the performance appraisal interview skills of managers on the videotapes, this seemed to be a plausible explanation for the purpose of the "research only" manipulation. After this brief introduction, subjects were asked to sign an informed consent form, which they all did.

After completing the consent forms, the subjects were asked to complete a brief biographical information form that was used to assess the acquaintance with the job variable. This questionnaire was a shorter form of the one used in Study 1; however, it still contained the same questions we used in Study 1 to assess the acquaintance with the job variable.

After completing this form, the subjects were briefed again on the study, its importance, and the "research only" purpose. The subjects then received either the BARS rating scales (Borman, 1978) or the SME-derived performance standards scale from Study 2, and a set of forms on which to make their ratings.

Depending on condition, the subjects either received a set of verbal instructions by the experimenter (EPVI) and were allowed to ask questions; were simply given a supplemental set of written instructions by the experimenter and were not allowed to ask questions (EPWI); or received the same set of instructions via a videotape of the experimenter and were not allowed to ask questions (VEVI). Thus, subjects were not permitted to ask questions in two of the three conditions, and had to rely on the written instructions or videotape explanation of the rating task.

The first data collection mode (EPVI) was similar to a research technician from AFHRL going to a field location to collect performance appraisal data. The second condition (EPWI) would be similar to sending written instructions to raters. The third condition (VEVI), using the videotape, would be similar to having an AFHRL technical person create a videotape for use in collecting performance appraisal data on Air Force personnel. In all conditions, the research importance of the study as a "$100,000 contract to SUNY-Albany for research only" was emphasized prior to starting the videotapes for data collection. The subjects were then shown each of the videotape sequences, and asked to rate the performance of the managers at the conclusion of each tape and complete the confidence ratings.

At the conclusion of the last videotape, the subjects were asked to complete a questionnaire which included the intervening variables, and the "quality of rating form" dependent variables. All subjects then received a lecture on how the results of the study in which they had participated were to be used by the AFHRL.

Results

Intervening Variables

Based on the analyses from Study 1, the scales with reduced items were included in this study, and were subjected to internal consistency reliability analyses. The alpha reliabilities based on 90 respondents were: (a) rater motivation, .72; (b) rater acceptance, .85; (c) rater trust, .48; and (d) rater confidence, .80. With the exception of rater trust, these reliabilities reached acceptable levels for research (Nunnally, 1978). Since the reliability for rater trust did not reach an acceptable level, it was dropped from further analyses.

MANOVA Results

A 3 (mode of data collection) by 2 (rating scale format) multivariate analysis of variance (MANOVA) was computed. The Hotellings test was significant ($p < .01$), indicating that there were significant effects for the independent variables. Given this result and the a priori hypotheses of this research, univariate ANOVAs were computed.

ANOVA Results

Since the hypotheses of this study dealt with relationships between the independent variables and both the intervening and dependent variables, separate ANOVAs were computed for these two sets of variables. In addition, the two quality of rating form dependent variables were analyzed separately.

Intervening Variables. There were no significant main or interaction effects between the independent and intervening variables.

Dependent Variables. Since we were able to calculate two scores for leniency, halo, correlational accuracy, and distance accuracy based on the Borman (1978) true scores and our SME-derived true scores from Study 3, we will annotate the results with either (Borman) or (SME) to indicate which score is being used.

There were no significant results for the different modes of data collection conditions.

The different rating scale formats produced several significant effects. As indicated in Table 10, there was significantly less leniency (SME) ($p < .05$, $w^2 = .04$) when the performance standards format was used.

There was a significant interaction for correlational accuracy (Borman) ($p < .05$, $w^2 = .05$) as seen in Figure 10. The higher the value, the greater the correlational accuracy. This interaction is primarily due to the reversal from the VEVI, performance standards condition to the EPWI, performance standards condition. It is interesting to note in Figure 10 that correlational accuracy is essentially the same for experimenter present (EPVI) with the BARS format condition and the performance standards format with only written instructions (EPWI) condition.

Table 10. **Means for Findings: Study 4**

Experimental treatment	Dependent variable: Range restriction[a]	Rating confidence[b]	Leniency[c]
BARS	1.54	3.77	.64
Performance Standards	1.62	3.56	.47

Note. Probability levels for the dependent variable findings are: Range Restriction ($p < .08$), Rating Confidence ($p < .07$), Leniency ($p < .05$).

[a]The higher the mean, the less the range restriction.

[b]The higher the mean, the higher the confidence.

[c]The higher the mean, the more leniency.

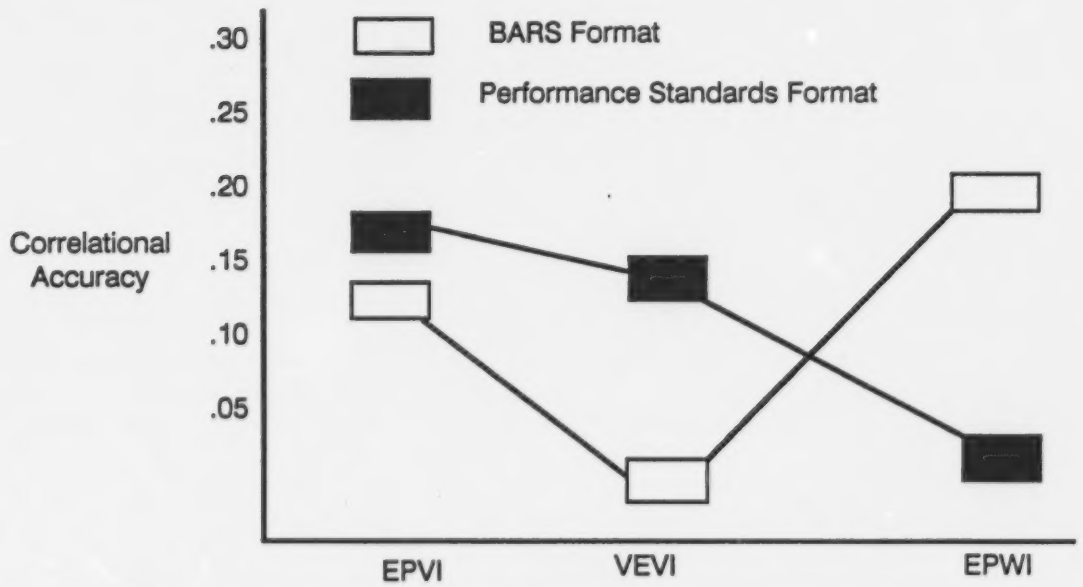

Figure 10. **Interaction for Correlational Accuracy (Borman): Study 4.**

Correlational Results

The correlational results are presented in Table 11. In terms of the acquaintance with the job variables (WE, PA, FB, SU), there are no consistent significant relationships with either the intervening or the dependent variables. The single significant relationship between supervisory experience and range restriction (RG) is most likely due to chance. These findings are more in agreement with the results from Study 1 than those from Study 2 with regard to the relationship between acquaintance with the job and performance measurement quality, and indicate that for this performance rating task, acquaintance with the job is of little importance.

Table 11. Correlation Results for Study 4

	Acquaintance with job			Quality of measurement								Intervening variables		
	PA	FB	SU	DA1	CA1	LN	HO	RG	CN	DA2	CA2	RM	RA	RC
WE	32	38	41	-02	04	01	-15	15	-08	-01	04	05	-05	-04
PA		79	71	09	-07	-04	-11	16	08	12	-12	11	01	-07
FB			81	10	-11	-05	-16	08	-02	-06	03	05	-09	-13
SU				16	-16	-08	-15	18	-01	06	-03	05	-04	-07
DA1					-90	35	-31	29	-07	-03	36	07	-04	09
CA1						-17	14	01	-10	21	-33	-01	02	-18
LN							00	-06	-02	06	12	-02	02	10
HO								-49	03	-13	-14	-14	16	07
RG									-10	53	02	28	01	00
CN										-05	06	23	43	-02
DA2											-77	24	08	01
CA2												-07	-03	12
RM													59	38
RA														63

Note. Decimals are omitted. For n = 90, correlations of .17 and .24 are significant at the .05 and .01 levels, respectively. WE = Work experience, PA = Performance appraisal experience, FB = Feedback experience, SU = Supervisory experience in general, DA1 = Distance accuracy (Borman), CA1 = Correlational accuracy (Borman), LN = Leniency, HO = Halo, RG = Range restriction, CN = Confidence in ratings, DA2 = Distance accuracy (SME), CA2 = Correlational accuracy (SME), RM = Rater motivation, RA = Rater acceptance, RC = Rater confidence.

The relationships between the intervening and dependent variables show some interesting findings. Two of the three intervening variables (rater trust was dropped due to low reliability) show a strong relationship with the confidence the subjects had in their job performance ratings. This is consistent with Study 1, and again, it is a self-report, self-report relationship. The other consistent set of relationships is between rater motivation and the three dependent variables of range, confidence, and distance accuracy (SME), thus supporting the hypothesized relationship (Figure 2) between this intervening variable and performance measurement quality. The final significant relationship is a negative one between correlational accuracy (Borman) and rater acceptance.

In terms of the accuracy dependent variables that have been emphasized in this research project, perhaps the most noteworthy finding of this study is the lack of significant relationships. Although disappointing, there are two possible reasons for this. First, it may be that, in terms of rating scale format, a BARS and a performance standards scale have the same effect on rating accuracy. This may be due to the fact that the rating task calls for a judgment decision, regardless of the rating scale format. It may be that if the comparision were made between the performance standards scale and one for which an observation decision had to be made (e.g., a Behaviorial Observation Scale [Latham & Wexley, 1977]), the performance standards scale would prove to be better.

Another possible reason for the lack of significant results may be that--given that the experimental conditions in this study were derived from the "best" that Studies 1 and 2 had to prescribe on the basis of accuracy results--there is no more incremental true variance that can be captured by the independent variables in this study. That is, it may be that the mode of data collection and the rating scale format are weak in terms of their relative effects on accuracy when purpose of rating, detail of instructions, and experimental technique are controlled to maximize rating accuracy.

When we examine the results for the other dependent variables, the performance standards scale appears slightly better than the BARS. There is less leniency in the ratings, and it is best with the written instructions condition, which also is the most cost effective. Further, there were no significant relationships with the two "quality of rating form" variables; this leads to the conclusion that the performance standards format, with its greater specificity, may be slightly better statistically, but probably not in terms of practical significance.

Based on the results of this study, it would clearly be inappropriate to spend the extra time converting an already existing BARS format to a performance standards format. However, it would probably be advisable, in the creation of a new performance rating scale during the BARS development, to emphasize the development of specific performance standards rather than focusing on behaviorial examples only, as is typically done. The greater specificity provided by focusing on performance standards during the BARS developmental stages should also be more defensible in case of litigation involving a performance appraisal system (Cascio & Bernardin, 1981).

The correlational results replicate some of the earlier results and also provide additional support for the model in Figure 2. The failure of the acquaintance with the job variable to demonstrate a relationship to either the intervening or dependent variables would indicate that, for this rating task, this variable should be dropped. However, as discussed in Study 1, the effect of this variable may change as a result of the type of job being evaluated, particularly with jobs involving highly technical tasks.

The positive relationships between rater motivation and the dependent variables replicate earlier results and provide additional support for the hypothesized relationships in Figure 2. It is clear these intervening variables must be included in future research in the JPMS project as well as related research in AFHRL.

In terms of the JPMS project and the validation of the ASVAB, these findings have some additional important implications. The fact that the mode of data collection had no impact on the dependent variables would indicate that it may be possible to use the least costly technique without sacrificing accuracy. It may be possible to collect performance ratings simply by using effective written rating scale instructions without having a technical person present. Of course, this presumes that both the rating scale and the instructions for completing it will be pre-tested as was done in this study.

Another important implication for the JPMS and the validation of the ASVAB, or other personnel programs, is that it is not necessary to develop rating scales with performance standards where good BARS scales already exist. However, it would probably be wise to place emphasis on performance standards in the development of new job performance rating scales for additional enlisted specialties in the JPMS project. These recommendations have important cost-savings implications, both in terms of maintaining performance measurement quality at the lowest cost and in terms of the defensibility of the rating scales.

REFERENCES

Aleamoni, L. M., & Hexner, P. Z. (1973). *The effect of different sets of instructions on student course and instructor evaluations.* Urbana: University of Illinois.

Aleamoni, L. M., & Hexner, P. Z. (1980). A review of the research on student evaluations and a report on the effects of different sets of instructions on student course and instructor evaluation. *Instructional Science*, 9, 67-84.

Alewine, T. (1982). Performance appraisal and performance standards. *Personnel Journal*, 61, 210-213.

Banks, C. G., & Murphy, K. R. (1985). Toward narrowing the research-practice gap in performance appraisal. *Personnel Psychology*, 38, 335-346.

Bartlett, C. J. (1983). What's the difference between valid and invalid halo? Forced-choice measurement without forcing a choice. *Journal of Applied Psychology*, 68, 218-226.

Bazerman, M. H., Beekun, R. I., & Schoorman, F. D. (1982). Performance evaluation in a dynamic context: A laboratory study of the impact of a prior commitment to the ratee. *Journal of Applied Psychology*, 67, 873-876.

Berkshire, J. R., & Highland, R. W. (1953). Forced choice performance rating: A methodological study. *Personnel Psychology*, 6, 356-378.

Bernardin, H. J., Orban, J. A., & Carlyle, J. J. (1981). Performance rating as a function of trust in appraisal and rater individual differences (pp. 311-315). *Proceedings of the 41st annual meeting of the Academy of Management*, San Diego, CA.

Borman, W. C. (1974). The rating of individuals in organizations: An alternative approach. *Organizational Behavior and Human Performance*, 12, 205-214.

Borman, W. C. (1978). Exploring the upper limits of reliability and validity in job performance ratings. *Journal of Applied Psychology*, 63, 135-144.

Borman, W. C., Hough, L., & Dunnette, M. (1976). *Performance ratings: An investigation of reliability, accuracy, and relationships between individual differences and rater error.* Minneapolis: Personnel Decisions, Inc.

Cascio, W. F., & Bernardin, H. J. (1981). Implications of performance appraisal litigation for personnel decisions. *Personnel Psychology*, 34, 211-226.

Centra, J. A. (1976). The influence of different directions on student ratings of instructors. *Journal of Educational Measurement*, 13(4), 266-282.

Cohen, J., & Cohen, P. (1975). *Applied regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum.

Delbecq, A. L., Van de Ven, A. H., & Gustafson, D. H. (1975). *Group techniques for program planning*. Glenview, IL: Scott, Foresman.

Department of Defense. (1984). *Armed Services Vocational Aptitude Battery (ASVAB) Information Pamphlet*, DOD 1304.12Z.

Driscoll, L. A., & Goodwin, W. L. (1979). The effects of varying information about use and disposition of results on university students' evaluations of faculty and courses. *American Educational Research Journal*, 16, 25-37

Duffy, J. F., & Kavanagh, M. J. (1983). Confounding the creation of social forces: Laboratory studies of negotiation behavior. *Journal of Conflict Resolution*, 27, 635-647.

Feldman, J. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology*, 66, 127-148.

Freeberg, N. (1968). Relevance of rater-ratee acquaintance in the validity and reliability of ratings. *Journal of Applied Psychology*, 53, 518-524.

Hakel, M. D. (1980). *An appraisal of performance appraisal: Sniping with a shotgun*. Discussant's comments presented at the 1st annual meeting of the Scientist-Practicioner Conference in Industrial/Organizational Psychology, Virginia Beach, VA.

Hedge, J. W., & Kavanagh, M. J. (1983). *Improving the accuracy of performance evaluations: A comparison of three methods of performance appraisal training*. Unpublished manuscript.

Jackson, S. E., & Zedeck, S. (1982) Explaining performance variability: Contributions of goal setting, task characteristics, and evaluative contexts. *Journal of Applied Psychology*, 67, 759-768.

Kane, J. (1980). *Alternative approaches to the control of systematic error in performance appraisals*. Paper presented at the 1st annual meeting of the Scientist-Practicioner Conference in Industrial/Organizational Psychology, Virginia Beach, VA.

Kavanagh, M. J. (1971). The content issue in performance appraisal: A review. *Personnel Psychology*, 24, 653-668.

Kavanagh, M. J., Borman, W. C., Hedge, J. W., & Gould, R. B. (1986, February). *Job performance measurement classification scheme for validation research in the military* (AFHRL-TR-85-51, AD-164 837). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Kavanagh, M. J., Hedge, J. W., DeBiasi, G. L., Miller, S., & Jones, R. (1983). *An empirically-based, multiple criteria approach to the design, development, and implementation of a performance measurement system*. Symposium presented at the annual meeting of the Academy of Management, Dallas, TX.

Kenny, D. A. (1979). *Correlation and causality*. New York: John Wiley.

Kirby, P. (1981). Part 1: A systematic approach to performance appraisal. *Management World*, 10(28), 16-17.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin*, 87, 72-107.

Latham, G. P., & Wexley, K. N. (1977). Behavioral observation scales for performance appraisal purposes. *Personnel Psychology*, 30, 255-268.

McIntyre, R. M., Smith, D. E., & Hassett, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. *Journal of Applied Psychology*, 69, 147-156.

Morano, R. (1979). An Rx for performance appraisal. *Personnel Journal*, 58, 306-307.

Murphy, K. R., & Balzer, W. K. (1986). Systematic distortions in memory-based behavior ratings and performance evaluations: Consequences for rating accuracy. *Journal of Applied Psychology*, 71, 39-44.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Ruddy, T. (1985). *Performance appraisal: A review of four training methods*. Unpublished master's thesis, Rensselaer Polytechnic Institute, Troy, NY.

Scott, W. E., & Hamner, W. C. (1975). The influence of variations in performance profiles on the performance evaluation process: An examination of the validity of the criterion. *Organizational Behavior and Human Performance*, 14, 360-370.

Sharon, A. T., & Bartlett, C. J. (1969). Effect of instructional conditions in producing leniency on two types of rating scales. *Personnel Psychology*, 22, 251-263.

Smith, D. E., Hassett, C. E., & McIntyre, R. M. (1982, April). *Using student ratings for administrative decision: Are ratings contaminated by perceived uses of the information*. Paper presented at the 23rd annual meeting of the Western Academy of Management, Colorado Springs, CO.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology*, 47, 149-155.

Sorcher, M., & Goldstein, A. P. (1972). A behavior modeling approach in training. *Personnel Administration*, 35(2), 35-41.

Stone, T. (1970, October). Sources of evaluator bias in performance appraisal. *Experimental Publication System*, 8, Ms. #290-12, 1-10.

Taylor, E. K., & Wherry, R. J. (1951). A study of leniency in two rating systems. *Personnel Psychology*, 4, 39-47.

Wherry, R. J., & Bartlett, C. J. (1982). The control of bias in ratings: A theory of ratings. *Personnel Psychology,* 35, 521-551.

Zedeck, S., & Cascio, W. (1982). Performance decision as a function of purpose of rating and training. *Journal of Applied Psychology*, 67, 752-758.

APPENDIX A: BIOGRAPHICAL QUESTIONNAIRE: STUDY 1

B BUS 584: Human Resources Management

Biographical Information

In order for us to match instructional material and style to the composition of this class, and for use with other exercises, we will be completing in class, please complete the following short questionnaire. Some of this information is already available in our files; however, it is easier for us if you enter on this form. Obviously, this information is also confidential, and we will treat it as such. Please return this form to the front of class after you have completed it. Thank you.

Name_____(please print)

Local Address_____

_____

_____

Local Phone _____

1. Sex _____

2. Age _____

3. Total years of full-time work experience (if any) _____

4. Undergraduate major _____

5. Total years experience as supervisor/manager (if any) _____

6. If you have completed performance appraisals for employees under your supervision, what was the approximate number you have done to date? _____

7. If you have provided feedback interviews on employees' performance, what was the approximate number to date?_____

8. Have you ever been a supervisor for engineers? If yes, for how many years? _____

9. Have you ever worked in any of the following activities? Please check all that apply.

_____ personnel selection

_____ performance appraisal

_____ attitude surveys

_____ labor negotiations

_____ job analysis

_____ college recruiting

_____ benefits administration

_____ human resources planning

_____ compensation

_____ EEO compliance

_____ OSHA programs

_____ job redesign

_____ job evaluation

_____ career development

_____ training programs

APPENDIX B: INSTRUCTIONS TO SUBJECTS: STUDY 1

Performance Appraisal Interviews: Script, Refer Back Condition

In this study, there are six different sequences involving the interaction of an engineering manager and his immediate supervisor. The engineering manager, Mr. Whipker, is the same person for all six sequences. He is an employee of the sponsoring organization from their Personnel Department. He was instructed to play the role of a disgruntled engineering manager in the performance appraisal interviews. There are six different managers in the six sequences. These are the individuals whose job performance is to be evaluated. That is, you are to evaluate how well they conduct this performance appraisal interview with this disgruntled engineering manager, Mr. Whipker. The interactions between "Mr. Whipker" and the six "supervisors" were tape recorded. A transcription of these tape recordings, prepared as a script of their meetings, is what you will be reading to make your ratings of the effectiveness of Whipker's manager in conducting the performance appraisal interview.

In making your ratings you will be using the rating forms that have been distributed to you. Please make all of your ratings on the forms that have been distributed following the instructions on the forms. Be certain to complete the ratings on all seven dimensions, and then your overall confidence in your ratings for each sequence at the bottom of the page. Be sure to complete all ratings for one sequence, and then come to the moderator to pick up the next sequence.

When completing your performance ratings, read through the entire typed "script" of the appraisal interview carefully. When you are making your ratings, you may refer back to this script as often as you like to help in your ratings. Feel free to page back through the script to help you make your performance ratings more accurate. When you finish your ratings on one script, return it to the moderator, and he/she will give you another script. If you have any questions, please ask the moderator in your session.

The performance appraisal interviews take place in the office of the Vice President for Engineering. The room contains a desk and chair, with another chair drawn up next to the desk. The V.P. for Engineering is seated at the desk when there is a knock at the door.

Performance Appraisal Interviews: Script, No Refer Back Condition

In this study, there are six different sequences involving the interaction of an engineering manager and his immediate supervisor. The engineering manager, Mr. Whipker, is the same person for all six sequences. He is an employee of the sponsoring organization from their Personnel Department. He was instructed to play the role of a disgruntled engineering manager in the performance appraisal interviews. There are six different managers in the six sequences. These are the individuals whose job performance is to be evaluated. That is, you are to evaluate how well they conduct this performance appraisal interview with this disgruntled engineering manager, Mr. Whipker. The interactions between "Mr. Whipker" and the six "supervisors" were tape recorded. A transcription of these tape recordings, prepared as a script of their meetings, is what you will be reading to make your ratings of the effectiveness of Whipker's manager in conducting the performance appraisal interview.

In making your ratings you will be using the rating forms that have been distributed to you. Please make all of your ratings on the forms that have been distributed following the instructions on the forms. Be certain to complete the ratings on all seven dimensions, and then your overall confidence in your ratings for each sequence at the bottom of the page. Be sure to complete all ratings for one sequence, and then come to the moderator to pick up the next sequence.

When completing your performance ratings, read through the entire typed "script" of the appraisal interview once, and then make your ratings. Do not refer back to the script after you have read it once. This is extremely important for this study. Again, read the script once carefully, but make your ratings without referring back to the script. When you finish your ratings on one script, return it to the moderator, and he/she will give you another script. If you have any questions, please ask the moderator in your session.

The performance appraisal interviews take place in the office of the Vice President for Engineering. The room contains a desk and chair, with another chair drawn up next to the desk. The V.P. for Engineering is seated at the desk when there is a knock at the door.

Performance Appraisal Interviews: Videotape Condition

In this study, there are six different videotaped sequences involving the interaction of an engineering manager and his immediate supervisor. The engineering manager, Mr. Whipker, is the same person for all six sequences. He is an employee of the sponsoring organization from their Personnel department. He was instructed to play the role of a disgruntled engineering manager in the performance appraisal interviews. There are six different managers in the six sequences. These are the individuals whose job performance is to be evaluated. That is, you are to evaluate how well they conduct this performance appraisal interview with this disgruntled engineering manager, Mr. Whipker.

In making your ratings, you will be using the rating forms that have been distributed to you. Please make all of your ratings on the forms that have been distributed, following the instructions on the forms. Be certain to complete the ratings on all seven dimensions, and then your overall confidence in your ratings for each sequence at the bottom of the page.

The performance appraisal interviews take place in the office of the Vice President for Engineering. The room contains a desk and chair, with another chair drawn up next to the desk. The V.P. for Engineering is seated at the desk when Mr. Whipker knocks at the door.

RATING FORM FOR USE WITH PERFORMANCE APPRAISAL INTERVIEW

NAME _____

SOCIAL SECURITY # _____

rating # _____

Instructions: Using the seven-point scale listed below, with seven as the highest rating and one as the lowest rating, circle the number that corresponds to your assessment of the employee being rated for each of the seven performance dimensions. After completing the ratings for each of the managers conducting the appraisal feedback interview , estimate how confident you feel that you have done an accurate assessment and fill in the appropriate response on the bottom of this form. You should complete a separate form for each employee you are rating.

high level performer			average level performer			low level performer
7	6	5	4	3	2	1

Dimension 1	Structuring the Interview	7	6	5	4	3	2	1
Dimension 2	Establishing Rapport	7	6	5	4	3	2	1
Dimension 3	Reacting to Stress	7	6	5	4	3	2	1
Dimension 4	Obtaining Information	7	6	5	4	3	2	1
Dimension 5	Resolving Conflict	7	6	5	4	3	2	1
Dimension 6	Developing the Employee	7	6	5	4	3	2	1
Dimension 7	Motivating the Employee	7	6	5	4	3	2	1

How confident are you about the ratings you just completed?

very highly confident	highly confident	moderately confident	slightly confident	not at all confident
5	4	3	2	1

MANAGER PERFORMANCE CATEGORIES

GUIDELINES FOR MAKING PERFORMANCE RATINGS

The next section of this booklet contains seven (7) Performance Categories describing effective, average, and ineffective performance on the job of manager in a problem-solving interview (Manager). The Performance Categories are designed to help you make accurate judgments about the performance of Managers on several important facets of this job.

The accompanying booklet entitled Manager Rating Scales should be used to record performance ratings you assign after referring closely to materials contained in the Performance Categories booklet. Now let's describe the features of the Performance Categories booklet and provide guidelines for proper use of the rating scales.

First, notice that each of the seven Performance Categories is labeled and defined carefully at the top of the page. In addition, directly below each category definition are three pairs of behaviorally oriented descriptors representing high level, average, and low level performance. Finally, below these descriptors are seven performance examples--specific behavioral examples of how Managers exhibiting various levels of effectiveness might perform on that category. The example numbered "7" demonstrates the highest level performance; the example numbered "1" demonstrates the lowest level.

Here is how you should use Performance Category information to rate a particular ratee. Referring first to Category A (Structuring and Controlling the Interview), read over the label and definition, and study the level descriptors and performance examples below. Then make a judgment about the performance level exhibited by the ratee by using both level descriptors and performance examples as benchmarks or guides. That is, evaluate the ratee by matching the level of performance he demonstrated with the level of performance indicated by level descriptors and performance examples. Remember, the ratee needs not exhibit performance exactly like the Manager depicted in one of the performance examples to rate him at that level. Instead, you should try to match the ratee's overall level of performance on that Performance Category with the level of performance represented by performance examples and level descriptors. When you feel you have "a match," record the appropriate rating in the Manager Rating Scales booklet. Follow this procedure for all seven Performance Categories.

THINGS TO GUARD AGAINST

Several sources of error can contribute to inaccuracies in your ratings. Here are a few suggestions for overcoming them:

1. Consider each Performance Category separately from all the rest. An almost universal error in ratings is called HALO ERROR. It occurs when the rater gives about the same ratings to a person on all aspects of performance. Usually this error occurs because a rater has not taken enough time to get clearly in mind what each separate category of performance refers to. Remember we are asking you to describe or evaluate each ratee on a number of <u>different</u> categories of performance. As you consider each of the persons you are rating, try to avoid getting into the habit of giving about the same rating to him on each Performance Category. Consider each category separately from all others. Be sure to rate all ratees in each category before going on to the next category.

2. Avoid using your own definitions for the various Performance Categories. A common reason for inaccurate ratings is that raters have different definitions of Performance Categories. This is why it is so very important for you to read the definitions, descriptors, and performance examples carefully. Avoid any previous impressions of what these things have meant to you. Base your ratings on the information provided in the Performance Category booklet.

3. Try to overcome the contrast effect which causes raters to underevaluate or overevaluate an individual because of the level of performance demonstrated by the ratee evaluated just before that individual. An individual tends to be underevaluated, for example, when he appears immediately after a high performer. Conversely, an individual tends to be overevaluated when he appears immediately after a poor performer. To overcome this rating error, attend carefully to the level descriptors and performance examples. Try not to compare one ratee with another; instead, judge each on his own merits, using the descriptors and performance examples as guides.

A. STRUCTURING AND CONTROLLING THE INTERVIEW

Clearly stating the purpose of the interview; maintaining control over the interview; displaying an organized and prepared approach to the interview versus not discussing the purpose of the interview displaying a confused approach; allowing Whipker to control the interview when inappropriate.

High Level Performance

- Outlines clearly the areas to be discussed and skillfully guides the discussion into those areas.
- Displays good preparation for the interview and effectively uses information about Whipker, his subordinates, etc. to conduct a well-planned interview.

EXAMPLES

What a high level performer might do:

7. At the start of the interview, this Baxter would be expected to outline clearly the areas he wished to discuss. He would then cover each of these areas by skillfully moving the discussion to a new topic whenever an impasse was reached.

6. This Baxter would be well prepared for the Whipker interview. He can be expected to display considerable knowledge about Whipker's projects and the qualifications of Whipker's subordinates.

78

Average Performance

- States the purpose of the interview but fails to cover some areas he intended to discuss.
- Appears prepared for the interview but at times is unable to control the interview or to guide it into areas planned for discussion.

PERFORMANCE EXAMPLES

What an average performer might do:

5. Can be expected to prepare some notes of some things to cover and occasionally refer to them during the interview.

4. Would expect this Baxter to state that the reason for their discussion was to talk about the communications failure which had occurred recently but that they could talk about other topics as well.

3. Can be expected to state that he has called Whipker in because he wants to get to know his people and to find out how they have been doing in their work.

Low Level Performance

- Fails to indicate the purpose of the interview and appears to be unfamiliar with the file information.
- Appears unprepared for the interview and is unable to control Whipker on the interview.

PERFORMANCE EXAMPLES

What a low level performer might do:

2. After offering a few pleasantries at the start of the interview, would expect this Baxter to be unsure about what to say next, and to remain silent and fidget with Whipker's personnel file.

1. Can expect this Baxter to seem unsure about where the interview is going and to allow Whipker to give him an ultimatum to either change the overtime rules or the delivery schedule on his contracts.

79

B. ESTABLISHING AND MAINTAINING RAPPORT

Setting an appropriate climate for the interview; opening the interview in a warm nonthreatening manner; being sensitive to Whipker versus setting a hostile or belligerent climate; being overly friendly or familiar during the interview; displaying insensitivity toward Whipker.

High Level Performance

. Draws Whipker out by projecting sincerity and warmth during the interview.

. Discusses Whipker's problems in a candid but nonthreatening and supportive way.

What a high level performer might do:

7. Would expect this Baxter to project considerable warmth and sincerity during the interview. He may be expected to discuss Whipker's job related problems candidly but in a nonthreatening manner, leaving Whipker with the feeling that his boss would support and help him do his job well.

6. Can be expected to draw Whipker out by talking about some of his problems as United Fund coordinator in his previous job, and then to ask Whipker about his own experience with the United Fund job

PERFORMANCE

Average Performance

. Displays some sincerity and warmth toward Whipker and indicates by his response to Whipker and his problems that he is reasonably sensitive to Whipker's work-related needs.

. Uses mechanical means to set Whipker to set Whipker at ease, i.e., offers coffee.

What an average performer might do:

5. Would be expected to begin the interview by saying that it was nice to talk to Whipker in an informal setting and that he hoped they would have a good working relationship.

4. Can expect this Baxter to greet Whipker cordially at the door and to offer him a chair.

3. Can be expected to begin the interview by slapping Whipker on the back and asking him how things are going on the job in such a manner that Whipker would feel somewhat uneasy.

PERFORMANCE

Low Level Performance

. Projects little feeling or sensitivity toward Whipker; makes no friendly gestures.

. Is confrontive and inappropriately blunt during the interview.

What a low level performer might do:

2. This Baxter would be expected to begin the interview somewhat abruptly by telling Whipker he had asked him in to talk about his (Whipker's) problems in the company.

1. This Baxter can be expected to tell Whipker, without any small talk, "I suppose we both know that you are here because we have been getting reports about your not being able to get along with people on the job."

PERFORMANCE

C. REACTING TO STRESS

Remaining calm and cool, even during Whipker's outbursts; apologizing when appropriate but not backing down or retreating unnecessarily; maintaining composure and perspective under fire versus reacting inappropriately to stress; becoming unreasonable, irate, or defensive in reaction to complaints; backing down inappropriately when confronted.

High Level Performance

- Remains calm during Whipker's outbursts and responds in a rational, problem-solving manner.
- Is firm but nondefensive in response to Whipker's verbal assaults; admits fault when aporopriate but maintains an effective, problem-solving orientation when interacting with Whipker.

What a high level performer might do:

7. Even though Whipker is at his assaultive best several times during the interview, this Baxter would still maintain his cool, his earnest voice, and his good eye-to-eye contact with Whipker.

6. If Whipker said that he wanted Baxter's job, this Baxter could be expected to be very calm and cool and to say, "Do you have any ideas as to why you didn't get it?"

P E R F O R M A N C E

89

82

Average Performance

- Maintains composure during most of the interview but may appear unsettled, self-conscious, or defensive in reaction to some of Whipker's outbursts.
- May become rattled when confronted but recovers quickly.

What an average performer might do:

5. If Whipker pressed him to explain why he didn't get Baxter's job, this Baxter would present his arguments in a low-key, logical manner.

4. Would expect this Baxter to become a bit rattled when Whipker blows off about the Valva incident, but to recover quickly and request more information about the run-in.

3. When Whipker complains about not receiving the memo regarding Tech Services, can expect this Baxter to say he had no idea what happened to the memo.

P E R F O R M A N C E

Low Level Performance

- Becomes aggressively authoritative with Whipker or becomes helplessly silent during Whipker's outbursts.
- Escalates conflict by reacting defensively to Whipker's outbursts or accusing Whipker of causing problems.

What a low level performer might do:

2. Can be expected to swallow hard and grab the sides of his chair when Whipker blows up about how he should have had Thompson's job.

1. Would expect this Baxter to respond to Whipker's belligerence by becoming belligerent himself and to state, "You got the memo as fast as anyone else--if you didn't receive the memo, it's your fault."

P E R F O R M A N C E

83

D. OBTAINING INFORMATION

Asking appropriate questions; probing effectively to ensure that meaningful topics and important issues are raised; seeking solid information versus glossing over problems and issues; asking inappropriate questions; failing to probe into Whipker's perception of problems.

High Level Performance

- Asks probing questions, ensuring that important topics are discussed.
- Through careful and effective questioning, is able to uncover substantive problems and issues.

What a high level performer might do:

7. By persistently, yet delicately probing Whipker's feelings, this Baxter would be able to determine that Whipker prefers technical to managerial work.

6. This Baxter can be expected to probe into several relevant areas without being overly offensive or upsetting Whipker.

PERFORMANCE EXAMPLES

Average Performance

- Asks general questions about Whipker's job and problems.
- Does some questioning and probing into important problems and job-related issues but generally fails to follow up effectively.

What an average performer might do:

5. Would expect this Baxter to ask Whipker how he (Whipker) liked his job and whether he had any problems.

4. Could be expected to ask Whipker why he left his former job.

3. This Baxter would be expected to do some probing but never to stick long with any subject that might be distasteful to him or to Whipker.

PERFORMANCE EXAMPLES

Low Level Performance

- Asks inappropriate or superficial questions which fail to confront important problems.
- Spends little or no time questioning Whipker about substantive problems or issues.

What a low level performer might do:

2. This Baxter may be expected, out of the blue, to ask Whipker to tell him about his feelings and emotions.

1. Would expect this Baxter to spend nearly the entire interview lecturing and cajoling Whipker and to make very little effort to obtain information from him.

PERFORMANCE EXAMPLES

E. RESOLVING CONFLICT

Moving effectively to reduce the conflict between Valva and Whipker, Whipker and subordinates, etc.; making appropriate commitments and setting realistic goals to ensure conflict resolution; providing good advice to Whipker about his relationships with Valva, subordinates, etc. versus discussing problems too bluntly or lecturing Whipker ineffectively regarding the resolution of conflict; failing to set goals or make commitments appropriate to effective conflict resolution; providing poor advice to Whipker about his relationships with Valva, subordinates, etc.

High Level Performance

. Effectively reduces conflict between Whipker and others by making appropriate and realistic commitments to help Whipker get along better in the department.

. Provides good advice about solving problems and about improving Whipker's poor relationships with his subordinates, Valva, etc.

What a high level performer might do:

7. Would expect this Baxter to explain patiently that disagreement between people such as the one between Whipker and Valva usually occur because they have different information. Can also be expected to urge Whipker to spend time with Valva to learn more about Valva's department in such a manner that Whipker would agree to do so.

6. This Baxter would offer to go with Whipker to see Valva for the purpose of working out solutions to the problems Whipker and Valva were having with each other.

Average Performance

. Puts forth some effort to reduce conflict between Whipker and others but usually does not commit himself to helping with this conflict resolution.

. Tends to smooth over problems and provide reasonably good advice to Whipker about conflict situations.

What an average performer might do:

5. Would expect this Baxter to tell Whipker very warmly that the disagreement with Valva was unfortunate but that he had confidence things would work out okay from now on.

4. When Whipker complains about Valva being incompetent, Baxter could be expected to mention that nobody can be perfect all the time and to urge Whipker to be more patient with him.

3. Can be expected to lecture at great length about treating others with respect and working harmoniously together.

Low Level Performance

. Lectures ineffectively or delivers inappropriate ultimatums to Whipker about improving his relationships with others or about changing his "attitude" toward people or problems.

. Fails to make commitments to help Whipker resolve problems or provides poor advice to Whipker about his relationships with Valva, subordinates, etc.

What a low level performer might do:

2. In response to Whipker's complaints about Valva, would expect this Baxter to state that Valva's department seemed to be running along pretty well. He would also be expected to argue at length about how competent Valva was.

1. This Baxter can be expected to tell Whipker in no uncertain terms that he does not tolerate dissension in his ranks and Whipker is not to mess up the Tech Services Department.

F. DEVELOPING WHIPKER

Offering to help Whipker develop professionally; displaying interest in Whipker's professional growth; specifying developmental needs and recommending sound developmental actions versus not offering to aid in Whipker's professional development; displaying little or no interest in Whipker's professional growth; failing to make developmental suggestions or providing poor advice regarding Whipker's professional development.

High Level Performance

- Displays considerable interest in Whipker's professional development and provides appropriate, high quality developmental suggestions.
- Makes commitments to help personally in Whipker's development.

What a high level performer might do:

PERFORMANCE

7. This Baxter can be expected to suggest that Whipker go through a series of job transfers three days a month so that Whipker can learn more about management and GCI. This Baxter can also be expected to say that he would be happy to review with Whipker on a regular basis what he (Whipker) had learned on these jobs.

6. Would expect this Baxter to tell Whipker that he should soften up a bit and temper his tough attitude without becoming a fake or changing his basic style. He would also be expected to offer to attend the Dale Carnegie Course with Whipker and to suggest that they both could benefit from it.

Average Performance

- Provides general developmental suggestions but usually fails to make a personal commitment to aid in Whipker's professional development.
- Shows moderate interest in Whipker's development; may direct Whipker to seek developmental suggestions elsewhere.

What an average performer might do:

PERFORMANCE

5. Can expect this Baxter to ask Whipker to head up the Project of the Year Committee, to offer help in organizing the committee, to offer and to talk with Whipker about problems as they arise.

4. Can be expected to offer Whipker help in his general development.

3. This Baxter would suggest that Whipker obtain a list of courses from the personnel department and take the ones he felt he needed.

Low Level Performance

- Expresses little or no interest in Whipker's professional development.
- Fails to offer developmental suggestions or provides poor advice regarding Whipker's professional growth and development.

What a low level performer might do:

PERFORMANCE

2. This Baxter could be expected to state that Whipker would have to work on his own to accomplish changes in his style.

1. If Whipker asked this Baxter for a list of things he could improve upon in order to get promoted, would expect him to be unable to come up with anything and to state that he didn't believe in training and development anyway.

G. MOTIVATING WHIPKER

Providing incentives for Whipker to stay at GCI and to perform effectively; making commitments or motivating Whipker to perform his job well, to remain with GCI, and to help GCI accomplish its objectives; supporting Whipker's excellent past performance versus providing little or no incentive for Whipker to stay at GCI and perform effectively; failing to make commitments encouraging Whipker's continued top performance; neglecting to express support of Whipker's excellent performance record.

High Level Performance

- A high level performer provides encouragement and appropriate incentives to persuade Whipker to stay with GCI and to perform effectively on his job.
- A high level performer uses appropriate compliments of Whipker's technical expertise and excellent past performance to motivate Whipker to meet the objectives of the department.

What a high level performer might do:

PERFORMANCE

7. This Baxter can be expected to tell Whipker he is "laying it on the line," and to state firmly that he and GCI need Whipker because of his impressive expertise and proven ability to get the job done. Can also expect him to ask Whipker's support in terms of continued top performance, to pledge in a sincere manner to do all he can to get Whipker more support in his present job, and to promise to seek out for Whipker more information about management and higher level technical job openings within GCI.

6. At the end of the interview, would expect this Baxter to reiterate the commitments he had made to Whipker with regard to inquiring about job openings in higher level technical positions within GCI and also to suggest that Whipker's excellent past performance and continued high level performance will increase his chances of getting such a job.

Average Performance

- An average performer compliments Whipker appropriately at times but is only moderately effective in using these compliments to encourage high performance, loyalty to GCI, etc.
- An average performer provides some incentives for Whipker to perform effectively and to stay at GCI, but generally makes few if any personal commitments to support Whipker in his job.

What an average performer might do:

PERFORMANCE

5. Would expect this Baxter to offer Whipker the United Fund job again in such a way that Whipker would agree to take it on, and then to say that he knew Whipker would do a good job because of his success in the past.

4. Throughout the interview, this Baxter can be expected to emphasize his desire to keep Whipker in the company.

3. Can be expected to tell Whipker he appears to be doing an adequate job in his department but that he could probably be doing better.

Low Level Performance

- A low level performer fails to express support for Whipker's past performance.
- A low level performer provides little or no incentive for Whipker to remain at GCI.

What a low level performer might do:

PERFORMANCE

2. This Baxter could be expected to tell Whipker to "keep plugging" on his job because GCI needs to increase its earnings.

1. After discussing Whipker's problems within GCI, this Baxter would suggest that he he (Whipker) leave the company since he was so dissatisfied

APPENDIX C: EXPERIMENTAL QUESTIONNAIRE: STUDY 1

Opinions on Performance Appraisal

Before beginning the items on this questionnaire, please answer the following question by circling the correct response:

What was the purpose of the performance ratings; that is, what are they to be used for?

a. part of a promotion decision
b. for research in validating a selection battery
c. for personal growth and development of the individuals
d. I don't know

In the remainder of this questionnaire, there are various items that ask your opinion about performance appraisal. The questions are concerned with your opinions about the performance ratings you have just made in this study. Answer each of the items in this questionnaire using the following scale. Place the number which corresponds to your answer in the blank space beside the question.

1 Not at all
2 To a little extent
3 To a moderate extent
4 To a great extent
5 To a very great extent

As with the other materials we have used in this study, please print your name where indicated. Again, we are only interested in matching your personal responses to this questionnaire with the other materials you have completed. Your individual responses to this questionnaire will remain totally confidential.

NAME_____

_____ 1. To what extent did you find the performance rating process boring?

_____ 2. To what extent do you feel other persons in this study really tried to follow the rules in completing their ratings?

_____ 3. To what extent do you believe that the true purpose of this study was the one explained in class?

_____ 4. To what extent was it very difficult for you to make the ratings?

_____ 5. To what extent are you confident we could use your ratings to determine merit pay raises for the employees depicted in the job situations?

_____ 6. To what extent are you confident that we could use your ratings as the performance measures for a $100,000 selection project?

_____ 7. To what extent do you feel you could defend your ratings to the others in your group who gave different ratings to the same employees in the job situations?

_____ 8. To what extent do you think other persons in this study gave higher ratings to help out the persons depicted in the job situations?

_____ 9. To what extent did you "inflate" your ratings to give the employees in the job situations a higher score?

_____ 10. To what extent did you care how accurate your ratings were in this study?

_____ 11. To what extent do you trust that the performance ratings you made are going to be used for the specific purpose described in the study?

_____ 12. To what extent do you feel other persons in this study really didn't care about making accurate ratings?

_____ 13. To what extent do you feel your ratings accurately captured the true performance of the people you rated?

_____ 14. To what extent were you uncertain as to which ratings to assign to specific employees?

_____ 15. To what extent did you understand how to complete the performance ratings?

_____ 16. Overall, to what extent did you feel confident about the ratings you made in this study?

_____ 17. To what extent do you believe that the data collected from you in this study is going to be used as described by the researchers?

_____ 18. To what extent would your closest friend describe you as a person who is overly concerned with accuracy in your work performance?

_____ 19. To what extent would you describe yourself as being overly concerned with details in other aspects of your life?

_____ 20. To what extent do you feel uncomfortable doing performance ratings that will have serious implications for the use of the results of this study?

_____ 21. To what extent were you clear as to the standards to use in making your ratings?

_____ 22. To what extent do you feel the performance ratings you completed are important to the sponsoring organization of this study?

_____ 23. To what extent do you feel ratings were the best way to measure the job performance of the employees?

_____ 24. To what extent did you understand what the ratings as described on the ratings form were trying to measure?

_____ 25. To what extent do you feel you were able to accurately distinguish between good and poor performers in the job situations?

_____ 26. To what extent were you uncomfortable giving negative ratings to the employees in the job situations?

_____ 27. To what extent do you feel the performance appraisals done in this study really measure the employees' true performance in the job situations depicted?

_____ 28. To what extent would you be willing to participate in another study of this kind later this semester?

_____ 29. To what extent do you feel other persons in this study were uncomfortable giving negative ratings to the employees in the job situations?

_____ 30. To what extent do you feel the ratings you made in this study accurately reflect the performance of the employees depicted in the job situations?

_____ 31. To what extent do you feel the results of this study will provide information to the sponsoring organization?

_____ 32. To what extent do you feel this was a useful study?

_____ 33. To what extent did you really make an "extra effort" to carefully pay attention to the job performance materials in order to make your ratings accurate?

_____ 34. To what extent did you enjoy completing the performance ratings in this study?

_____ 35. To what extent do you feel the ratings you completed in this study are going to be useful and worthwhile for this research?

_____ 36. To what extent do you feel the results of this study will be useful for application in real organizations?

_____ 37. Given the circumstances of this study, to what extent were you very willing to complete the ratings?

_____ 38. To what extent do you feel the rating form used in this study is a good one?

_____ 39. To what extent did the rating form enable you to evaluate the performance of the employees in the job situations fairly?

_____ 40. To what extent did you feel it was important for you to make accurate ratings in this study?

_____ 41. Based on your experience in this study, how important is it to you to make any performance ratings you do in the future as accurate as you can?

_____ 42. To what extent did the rating form used in this study enable you to make accurate ratings?

_____ 43. To what extent are you satisfied you made the most accurate ratings you could in this study?

Thank you very much for your help in completing this study. If you are willing to participate in another performance rating study similar to this one later this semester, please print your name, local address and phone number below. We will be in touch with you sometime in mid-April.

Name:_____

Address:_____

_____

Phone:_____

APPENDIX D: INSTRUCTIONS TO SUBJECT MATTER EXPERTS

May 15, 1985

&title& &fname& &lname&
&position/o&
&company&
&street/o&
&city&

Dear &title& &lname&:

I am involved in a long-term research project funded by the U. S. Air Force Human Resources Laboratory. In general, this project is concerned with developing the best and most accurate rating system for the evaluation of individual job performance. As one part of this research project, I need qualified persons to serve as Subject- Matter Experts (SMEs) in the field of performance appraisal. A brief description of the project with a definition of SMEs and the level of participation required is enclosed.

Since the required expertise for SMEs is Human Resources Managers, I am seeking participants from the HR community in the Capitol District. Twelve SMEs will be needed for this project, and it is quite appropriate to have more than one person from a participating company as long as each has had supervisory experience in conducting performance appraisal feedback interviews.

The timetable for this project is as follows. Task 1 will be completed on June 18, 1985 at a meeting on campus from 8:30 to 11:00 a. m. Study participants will receive a copy of the videotapes (VHS) and the scripts for the tapes. The SMEs will review the tapes and scripts, on their own, and rate the performance of the managers. On the following Tuesday, June 25, the SMEs will meet again from 8:30 to 11:00 to complete task three.

For their participation, SMEs will receive an honorarium of $50.00. I realize this is a small amount for the level of participation; however, participants will also have access to the materials used in this study for their own use, perhaps in supervisory training programs. Furthermore, all participants will receive periodic reports on the progress of this project for the next several years.

I would like to secure all participants for this project by June 1, 1985. Therefore, would you please let me know by letter of any persons in your organization willing to participate. We will be in touch with them directly to finalize the arrangements.

If you have any questions about this project, please call me at 457-8515 (O) or 439-1313 (H). Thank you for your willingness to consider helping me with this project.

Sincerely,

MJK/ah
Enc.

Michael J. Kavanagh, Ph. D.
Professor of Management and
Project Director

PERFORMANCE APPRAISAL ACCURACY PROJECT

Purpose

The general purpose of this project is to improve the accuracy of performance appraisals done in the Armed Forces. Specifically, this applied research project is examining aspects of the performance rating process that can affect the accuracy of performance ratings made by supervisors. We are concerned about eliminating errors in this rating process. To do this, we have designed a series of four studies using videotapes of actors playing the roles of a manager and one of his subordinate managers during a performance appraisal review session. There are six different sequences in which the subordinate manager remains the same, and his manager is different. The manager's role was designed to be different in the six tapes such that some managers are more effective than others. The different managers are rated on their performance using a standard rating form developed for this research project.

Participation Needed

Since the videotapes were developed some time ago, they need better calibration to be useful in the research studies. Thus, this study is designed to use Subject-Matter Experts (SMEs) to examine both the videotapes and the rating form to determine their quality. Specifically, the SMEs will be involved in the following three tasks:

1. Developing specific performance standards for the rating forms used to rate the videotaped performance of the manager. This will be done as a group and should take two to three hours.

2. Rating how well the six managers handle the performance appraisal interview by viewing the tapes. This can be done on your own, and will take approximately four hours. You will also have the original scripts the actors used for their parts.

3. Reaching consensus among the group of SMEs on the "correct" ratings for each videotaped sequence. This will be done together, and should take approximately two hours.

Subject Matter Experts

The SMEs being sought for this project are Human Resources Managers who have at least three years experience at completing performance appraisals and conducting performance appraisal interviews. HR Managers are wanted because of their expertise, and they may be able to use the materials from the project in their own organizations. Participants will receive reports on the progress of this research in terms of specific recommendations for improvements in performance appraisal accuracy.

GENERAL PERFORMANCE STANDARDS

The following criteria are to be used in establishing performance ratings:

Excellent: Accomplishments and results consistently exceed the normal and expected level of work. The staff member makes significant contributions to the objective of the department; rarely needs assistance in completing assignments;demonstrates creativity and ingenuity in solving problems. Achievements are clearly apparent to all.

Good: Accomplishments and results generally exceed the expected level of work. The staff member meets all objectives and goals; gives extra effort to get the job accomplished; needs a minimum of supervision.

Satisfactory: Accomplishments and results generally meet the expected level of work. The staff member is steady and dependable in performance of duties; is representative of the solid, dependable conscientious worker who forms the nucleus of any department.

Less than Satisfactory: Accomplishments and results are generally below the expected level of work, and are at best minimally acceptable. Further counseling, training, and experience appear necessary to raise performance to a satisfactory level.

Unsatisfactory: Accomplishments and results do not meet the expected level of work. The staff member is unwilling or unable to meet work expectations. The work is unacceptable.

APPENDIX E: LOW LEVEL OF DETAIL INSTRUCTIONS

Performance Appraisal Interviews

In this study, there are five different videotaped sequences involving the interaction of an engineering manager and his immediate supervisor. The engineering manager, Mr. Whipker, is the same person for all five sequences. He is an employee of the sponsoring organization from their Personnel department. He was instructed to play the role of a disgruntled engineering manager in the performance appraisal interviews. There are five different managers in the five sequences. These are the individuals whose job performance is to be evaluated. That is, you are to evaluate how well they conduct this performance appraisal interview with this disgruntled engineering manager, Mr. Whipker.

In making your ratings, you will be using the rating forms that have been distributed to you. These are stapled together, but please print your name and social security number where indicated on each form. In the space for rating #, write 1 for the first script, 2 for the second, and so on through all five scripts. In making your ratings from 1 to 7 on the dimensions, you should refer to the detailed decriptions of the performance dimensions distributed to you with the rating forms. Please make all of your ratings on the forms that have been distributed, following the instructions on the forms. Be certain to complete the ratings on all seven dimensions, and then your overall confidence in your ratings for each sequence at the bottom of the page.

The performance appraisal interviews take place in the office of the Vice President for Engineering. The room contains a desk and chair, with another chair drawn up next to the desk. The V.P. for Engineering is seated at the desk when Mr. Whipker knocks at the door.

APPENDIX F: MODERATE LEVEL OF DETAIL INSTRUCTIONS

GUIDELINES FOR MAKING PERFORMANCE RATINGS

The next section of this booklet contains seven (7) Performance Categories describing effective, average, and ineffective performance on the job of manager in a problem-solving interview (Manager). The Performance Categories are designed to help you make accurate judgments about the performance of Managers on several important facets of this job.

The accompanying booklet entitled Manager Rating Scales should be used to record performance ratings you assign after referring closely to materials contained in the Performance Categories booklet. Now let's describe the features of the Performance Categories booklet and provide guidelines for proper use of the rating scales.

First, notice that each of the seven Performance Categories is labeled and defined carefully at the top of the page. In addition, directly below each category definition are three pairs of behaviorally oriented descriptors representing high level, average, and low level performance. Finally, below these descriptors are seven performance examples--specific behavioral examples of how Managers exhibiting various levels of effectiveness might perform on that category. The example numbered "7" demonstrates the highest level performance; the example numbered "1" demonstrates the lowest level.

Here is how you should use Performance Category information to rate a particular ratee. Referring first to Category A (Structuring and Controlling the Interview), read over the label and definition, and study the level descriptors and performance examples below. Then make a judgment about the performance level exhibited by the ratee by using both level descriptors and performance examples as benchmarks or guides. That is, evaluate the ratee by matching the level of performance he demonstrated with the level of performance indicated by level descriptors and performance examples. Remember, the ratee needs not exhibit performance exactly like the Manager depicted in one of the performance examples to rate him at that level. Instead, you should try to match the ratee's overall level of performance on that Performance Category with the level of performance represented by performance examples and level descriptors. When you feel you have "a match," record the appropriate rating in the Manager Rating Scales booklet. Follow this procedure for all seven Performance Categories.

THINGS TO GUARD AGAINST

Several sources of error can contribute to inaccuracies in your ratings. Here are a few suggestions for overcoming them:

1. Consider each Performance Category separately from all the rest. An almost universal error in ratings is called HALO ERROR. It occurs when the rater gives about the same ratings to a person on all aspects of performance. Usually this error occurs because a rater has not taken enough time to get clearly in mind what each separate category of performance refers to. Remember we are asking you to describe or evaluate each ratee on a number of <u>different</u> categories of performance. As you consider each of the persons you are rating, try to avoid getting into the habit of giving about the same rating to him on each Performance Category. Consider each category separately from all others. Be sure to rate all ratees in each category before going on to the next category.

2. Avoid using your own definitions for the various Performance Categories. A common reason for inaccurate ratings is that raters have different definitions of Performance Categories. This is why it is so very important for you to read the definitions, descriptors, and performance examples carefully. Avoid any previous impressions of what these things have meant to you. Base your ratings on the information provided in the Performance Category booklet.

3. Try to overcome the contrast effect which causes raters to underevaluate or overevaluate an individual because of the level of performance demonstrated by the ratee evaluated just before that individual. An individual tends to be underevaluated, for example, when he appears immediately after a high performer. Conversely, an individual tends to be overevaluated when he appears immediately after a poor performer. To overcome this rating error, attend carefully to the level descriptors and performance examples. Try not to compare one ratee with another; instead, judge each on his own merits, using the descriptors and performance examples as guides.

APPENDIX G: HIGH LEVEL OF DETAIL INSTRUCTIONS

DETAILED INSTRUCTIONS FOR RATING FORM

Completing ratings of job performance is a difficult task. In this rating form, we have tried to make this task easier. These instructions will take you step-by-step through the correct procedure to use in rating the performance of the managers you will see on the videotapes.

The attached rating form is called a behaviorally anchored rating form because the rating scale numbers for the Performance Dimensions have specific examples of the behavior corresponding to that level of performance. Each of the seven Performance Dimensions is labeled and defined carefully at the top of the page. In addition, directly below each dimension definition are three pairs of behaviorally oriented descriptors representing the high level, average, and low level performance. Finally, below these descriptors are seven performance examples -- specific behavioral examples of how Managers exhibiting various levels of effectiveness might perform on that dimension. The example numbered "7" demonstrates the highest level of performance; the example numbered "1" demonstrates the lowest level. Take a look at the seven performance dimensions now before you continue with these instructions.

HOW TO USE THE RATING SCALE

The best way to make the most accurate ratings in this study is to follow the following steps.

Step 1. After viewing one tape involving the interaction between Baxter and Whipker, start by reading the definition for the first performance dimension -- "Structuring and Controlling the Interview."

Step 2. First decide which of three general performance levels -- high, average, or low -- best describes the overall performance that Baxter exhibited on this performance dimension.

Step 3. Now go immediately below the general performance level you have chosen, and try to determine which specific performance level best fits Baxter's performance in the script you have just read. In making this specific judgment, try to recall specific examples of Baxter's performance during the performance interview. Remember, you can refer back to the script to check for these specific examples. When you have decided on the

specific performance level, write the number corresponding to your rating on the rating form.

Step 4. Follow the above three steps for the remaining performance dimensions.

THINGS TO GUARD AGAINST

Several sources of errors can contribute to inaccuracies in your ratings. Here are a few suggestions for overcoming them.

1. Consider each Performance Dimension separately from all the rest. An almost universal error in ratings is called HALO ERROR. It occurs when the rater gives about the same ratings to a person on all aspects of performance. Usually this occurs because a rater has not taken enough time to get clearly in mind what each separate dimension of performance refers to. Remember we are asking you to describe or evaluate each ratee on a number of different dimensions of performance. As you consider each of the persons you are rating, try to avoid getting into the habit of giving about the same rating to him on each Performance Dimension. Consider each dimension separately from all others.

2. Avoid using your own definition for the various Performance Dimensions. A common reason for inaccurate ratings is that raters have different definitions of Performance Dimensions. This is why it is so very important for you to read the definitions, decriptors, and performance examples carefully. Avoid any previous impressions of what these things have meant to you. Base your ratings on the information provided in the Performance Dimensions rating scale.

3. Try to overcome the CONTRAST EFFECT which causes raters to under-evaluate or over-evaluate an individual because of the level of performance demonstrated by the ratee evaluated just before that individual. An individual tends to be under-evaluated, for example, when he appears immediately after a high performer. Conversely, an individual tends to be over-evaluated when he appears immediately after a poor performer. To overcome this rating error, attend carefully to the level descriptors and the performance examples. Try not to compare one ratee with another; instead, judge each on his own merits, using the descriptors and performance examples as guides.

APPENDIX H: BIOGRAPHICAL INFORMATION AND QUESTIONNAIRE

Biographical Information

Please complete the following short questionnaire. The information will be used in conjunction with the experimental materials you complete for this study. Obviously, this information is confidential, and we will treat it as such. Please return this form to the front of the room after you have completed it. Thank you.

Name_____(please print)

Social Security # _____

1. Sex _____
2. Age _____
3. Total years of full-time work experience (if any) _____
4. Total years experience as supervisor/manager (if any) _____
5. If you have completed performance appraisals for employees under your supervision, what is the approximate number you have done to date? _____
6. If you have provided feedback interviews on employees' performance, what is the approximate number to date?_____
7. Have you ever been a supervisor for engineers? If yes, for how many years? _____

Opinions on Performance Appraisal

Before beginning the items on this questionnaire, please answer the following question by circling the correct response:

What was the purpose of the performance ratings; that is, what are they to be used for?

a. part of a promotion decision
b. for research in validating a selection battery
c. for personal growth and development of the individuals
d. I don't know

In the remainder of this questionnaire, there are various items that ask your opinion about performance appraisal. The questions are concerned with your opinions about the performance ratings you have just made in this study. Answer each of the items in this questionnaire using the following scale. Place the number which corresponds to your answer in the blank space beside the question.

1 Not at all
2 To a little extent
3 To a moderate extent
4 To a great extent
5 To a very great extent

_____ 1. To what extent do you believe that the true purpose of this study was the one explained by the researcher?

_____ 2. To what extent are you confident we could use your ratings to evaluate test scores (validate) on the employees depicted in the job situations?

_____ 3. To what extent do you feel you could defend your ratings to the others in your group who gave different ratings to the same employees in the job situations?

Quality 4. To what extent did the directions for using the rating scale help you to use it properly?

_____ 5. To what extent do you think other persons in this study gave higher ratings to help out the persons depicted in the job situations?

1 Not at all
2 To a little extent
3 To a moderate extent
4 To a great extent
5 To a very great extent

_____ 6. To what extent did you "inflate" your ratings to give the employees in the job situations a higher score?

_____ 7. To what extent do you trust that the performance ratings you made are going to be used for the specific purpose described in the study?

_____ 8. To what extent do you feel your ratings accurately captured the true performance of the people you rated?

_____ 9. Overall, to what extent did you feel confident about the ratings you made in this study?

_____ 10. To what extent would your closest friend describe you as a person who is overly concerned with accuracy in your work performance?

Understand 11. To what extent were you clear as to the standards to use in making your ratings?

Understand 12. To what extent did you understand what the ratings as described on the ratings form were trying to measure?

_____ 13. To what extent were you uncomfortable giving negative ratings to the employees in the job situations?

_____ 14. To what extent would you be willing to participate in another study of this kind in the future?

Understand 15. To what extent did you understand how to complete the performance ratings?

_____ 16. To what extent do you feel the results of this study will provide information to the sponsoring organization?

_____ 17. To what extent did you really make an "extra effort" to carefully pay attention to the job performance materials in order to make your ratings accurate?

1 Not at all
2 To a little extent
3 To a moderate extent
4 To a great extent
5 To a very great extent

_____ 18. To what extent would you describe yourself as being overly concerned with details in other aspects of your life?

_____ 19. To what extent did you enjoy completing the performance ratings in this study?

_____ 20. To what extent do you feel the ratings you completed in this study are going to be useful and worthwhile for this research?

_____ 21. To what extent do you feel you were able to accurately distinguish between good and poor performers in the job situations?

_____ 22. To what extent do you feel the results of this study will be useful for application in real organizations?

_____ 23. To what extent are you confident that we could use your ratings as the performance measures for a $100,000 selection project?

_____ 24. To what extent do you feel this was a useful study?

_____ 25. Given the circumstances of this study, to what extent were you very willing to complete the ratings?

Quality 26. To what extent do you feel the rating form used in this study is a good one?

_____ 27. To what extent did you feel it was important for you to make accurate ratings in this study?

_____ 28. Based on your experience in this study, how important is it to you to make any performance ratings you do in the future as accurate as you can?

1 Not at all
2 To a little extent
3 To a moderate extent
4 To a great extent
5 To a very great extent

Quality 29. To what extent did the rating form used in this study enable you to make accurate ratings?

_____ 30. To what extent are you satisfied you made the most accurate ratings you could in this study?

Quality 31. To what extent did the rating form enable you to evaluate the performance of the employees in the job situations fairly?

Understand 32. To what extent do you feel the written instructions with the rating form were completely clear?

Thank you very much for your help in completing this study. If you are willing to participate in another performance rating study similar to this one in the future, please print your name, local address, and phone number below. We will be in touch with you sometime during the Spring semester.

Name: _____

Address: _____

_____

Phone: _____

APPENDIX I: PERFORMANCE STANDARDS RATING FORM

DIMENSION A: ORGANIZING AND MANAGING THE INTERVIEW

Preparing a plan for the interview; communicating the purpose of the interview to the employee beforehand; keeping the interview "on track," while remaining flexible enough to depart from the plan if need be; VERSUS not discussing the purpose of the interview; displaying a confused approach; allowing Whipker to control the interview.

HIGH LEVEL PERFORMANCE

*Outlines clearly the areas to be discussed and skillfully guides the discussion in to those areas.
*Anticipates potential problems, shows flexibility in dealing with unexpected issues, and returns to the agenda.
*Displays good preparation for the interview, and initiates problem-solving.

What a high level performer might do:

7. At the start of the interview, this Baxter would be expected to outline clearly the areas he wished to discuss. Baxter would display flexibility in dealing with issues outside his agenda, but would skillfully lead the discussion back to his plan.

6. This Baxter would be well-prepared for the Whipker interview, and would have communicated the interview's purpose to Whipker ahead of time. Baxter would lead the interview in such a way that all his agenda items would be discussed.

AVERAGE LEVEL PERFORMANCE

*States the purpose of the interview, and attempts to cover all items on his agenda.
*Appears prepared for the interview, but at times is unable to keep the discussion "on track".
*Anticipates some potential sources of conflict.

What an average performer might do:

5. Can be expected to prepare notes of some subjects to discuss, and occasionally refer to them during the interview. Makes note of additional issues that are brought up, but returns to the topic at hand.

4. Would expect this Baxter to state that the reason for their discussion was to talk about the communications failure that occurred recently, but that they could talk about other topics as well. Has made a list of other topics he wants to touch on.

3. Can be expected to state that he has called Whipker in because he wants to get to know his people and find out how they have been doing in their work.

LOW LEVEL PERFORMANCE

*Fails to indicate the purpose of the interview, and appears to be unfamiliar with the information in the personnel folder.
*Is unaware of potential problems.
*Appears unprepared for the interview, and is unable to manage the direction of the interview.

What a low level performer might do:

2. After offering a few pleasantries at the start of the interview, would expect this Baxter to be unsure about what to say next, and to remain silent and fidget with Whipker's personnel file.

1. Can expect this Baxter to seem unsure about where the interview is going and to allow Whipker to lead the interview.

DIMENSION B: ESTABLISHING AND MAINTAINING RAPPORT

Opening the interview in a warm, nonthreatening manner; maintaining the employee's self-esteem; listening and being sensitive to Whipker, and enlisting his help in problem-solving VERSUS Being overly friendly or familiar during the interview; setting a hostile or belligerent climate; displaying insensitivity to Whipker.

HIGH LEVEL PERFORMANCE

*Draws Whipker out by projecting sincerity and warmth during the interview. Pays attention to Whipker's concerns and viewpoints.
*Listens to Whipker and responds with empathy.
*Discusses Whipker's problems in a candid but nonthreatening and supportive way.

What a high level performer might do:

7. Would expect this Baxter to project warmth and sincerity during the interview. He may be expected to enlist Whipker's help in solving job-related problems. He leaves Whipker with the feeling that his boss would support him and help him do his job well.

6. Can be expected to draw Whipker out about job-related problems, and give close attention to his answers. Baxter might share with Whipker some of his own previous experiences in a prior job.

AVERAGE PERFORMANCE

*Displays some sincerity and warmth toward Whipker and indicates by his response to Whipker and his problems that he is reasonably sensitive to Whipker's work-related needs.
*Uses mechanical means to set Whipker at ease, i.e., offers coffee.
*Greets Whipker warmly and engages him in a moment of "small talk" before beginning the interview.

What an average performer might do:

5. Would be expected to begin the interview by saying that it was nice to talk to Whipker in an informal setting and that he hoped they would have a good working relationship. Fails to listen consistently during the interview.

4. Can expect this Baxter to greet Whipker cordially at the door and to offer him a chair.

3. This Baxter would be expected to begin the interview somewhat abruptly by telling Whipker he had asked him in to talk about his (Whipker's) problems in the company.

LOW LEVEL PERFORMANCE

*Projects little feeling or sensitivity toward Whipker; makes no friendly gestures.
*Attempts to be friendly, but appears phony or insincere.
*Is confrontive and inappropriately blunt during the interview. Makes no attempt to get Whipker's views on any issue.

What a low level performer might do:

2. Can be expected to begin the interview by slapping Whipker on the back and asking him how things are going on the job in such a manner that Whipker would feel somewhat uneasy.

1. This Baxter can be expected to tell Whipker, without any small talk, "I suppose we both know that you are here because we have been reports about your not being able to get along with people on the job."

DIMENSION C: REACTING TO STRESS IN THE INTERVIEW

Keeping the discussion job-related; accepting responsibility for a mistake, but not backing down or retreating unnecessarily; maintaining composure and perspective under fire VERSUS becoming unreasonable, irate, or defensive in reaction to complaints; backing down inappropriately when confronted.

HIGH LEVEL PERFORMANCE

*Remains calm during Whipker's outbursts and responds in a rational, problem-solving manner.
*Keeps the discussion job-related.
*Is firm but nondefensive in response to Whipker's verbal assaults; accepts responsibility for errors but maintains an effective, problem-solving approach when interacting with Whipker.

What a high level performer might do:

7. This Baxter maintains his cool, his earnest voice, and his good eye-to-eye contact. If the situation appears too far gone, this Baxter might suggest that they end their meeting, cool down, and resume their discussion at a later time.

6. If Whipker said that he wanted Baxter's job, this Baxter could be expected to be very calm and to defuse the situation, and move on to another topic.

AVERAGE LEVEL PERFORMANCE

*Maintains composure during most of the interview but may appear unsettled, self-conscious or defensive in reaction to some of Whipker's outbursts
*May become rattled when confronted but recovers quickly.
*Uses the "shared problem" approach rather than reacting defensively.

What an average performer might do:

5. If Whipker pressed him to explain why he didn't get Baxter's job, this Baxter would present his arguments in a logical, low-key manner.

4. Would expect this Baxter to become a bit rattled when Whipker blows off about the Valve incident, but to recover quickly and request more information about the background of the conflict.

LOW LEVEL PERFORMANCE

*Allows his emotions to get the better of him, and worsens a bad situation.
*Becomes aggressively authoritative with Whipker or becomes helplessly silent during Whipker's outbursts.
*Escalates conflict by reacting defensively to Whipker's complaints or accusing Whipker of causing problems.

What a low level performer might do:

3. When Whipker complains about not receiving the memo regarding Tech Services, can expect this Baxter to say he had no idea what happened to the memo.

2. Becomes visibly upset and seems intimidated by Whipker's outbursts.

1. Would expect this Baxter to respond to Whipker's belligerence by becoming belligerent himself.

DIMENSION D: OBTAINING INFORMATION

Having good preliminary information before interview; asking appropriate questions and listening carefully to the answers; probing effectively to ensure that important issues are raised; VERSUS glossing over problems and issues; asking inappropriate questions; failing to listen to Whipker's answers or clarify ambiguous answers.

HIGH LEVEL PERFORMANCE

*Asks probing questions, ensuring that important topics are discussed.
*Through careful questioning and effective listening, is able to uncover substantive issues and problems.
*Follows up on questions that are answered incompletely, so that he gets enough information to do his job.

What a high level performer might do:

7. Asks questions with the goal of gaining factual information, and where appropriate, Whipker's opinions on problems in his department.

6. Uses questions to confirm information from other sources, and is sure he understands the answer before moving to another topic.

AVERAGE PERFORMANCE

*Does some questioning and probing into important problems and job-related issues, but generally fails to follow up effectively.
*Asks general questions about Whipker's job and problems.

What an average performer might do:

5. This Baxter can be expected to probe into several relevant areas without upsetting Whipker. This Baxter does not stick with an area that might be distasteful to him or to Whipker.

4. Would expect this Baxter to ask Whipker how he liked his job, and whether he had any problems.

3. Could be expected to ask Whipker why he left his former job.

LOW LEVEL PERFORMANCE

*Asks inappropriate or superficial questions which fail to confront important problems.
*Spends little or no time questioning Whipker about substantive issues or problems.

What a low level performer might do:

2. This Baxter may be expected, out of the blue, to ask Whipker to tell him about his feelings and emotions.

1. Would expect this Baxter to spend nearly the entire interview lecturing and cajoling Whipker, and to make very little effort to obtain information from him.

DIMENSION E: DEALING WITH INTERPERSONAL CONFLICT

Providing good advice to Whipker about his relationships with Valva, his subordinates, etc.; making appropriate commitments and setting realistic goals to help Whipker develop and use his own conflict resolution skills; moving effectively to reduce the conflict between himself and Whipker about the recent promotion VERSUS discussing problems too bluntly or lecturing Whipker about the resolution of the conflict with Valva; blaming Whipker for conflicts; glossing over the conflicts that currently exist between Whipker and Valva, and Whipker and Baxter.

HIGH LEVEL PERFORMANCE

*Anticipates potential areas of conflict, enlists Whipker's help in resolving conflict, and commits the time necessary to monitor Whipker's efforts.
*Effectively reduces conflict between Whipker and others by making appropriate and realistic commitments to help Whipker get along better in the department.

What a high level performer might do:

7. Would expect this Baxter to convince Whipker that his problems with Valva are jointly caused and must be jointly solved. Can also be expected to help Whipker develop a plan for approaching Valva to begin solving their interpersonal conflicts.

6. This Baxter suggests that Whipker make a list of his needs from the Tech Services department, and then go to discuss the list with Valva.

AVERAGE LEVEL PERFORMANCE

*Provides good advice about solving problems and about improving Whipker's poor relationships on the job. Also tries to enlist Whipker's support in developing solutions to the conflicts.
*Tends to smooth over problems, but provides good advice to Whipker about conflict situations.
*Puts forth some effort to reduce conflict between Whipker and others but does not commit himself to helping with this conflict resolution.

What an average performer might do:

5. This Baxter would offer to go with Whipker to see Valva for the purpose of working out solutions to the problems Whipker and Valva were having with each other.

4. When Whipker complains about Valva being incompetent, Baxter could be expected to state that nobody is perfect, and to urge Whipker to be more patient with Valva. This Baxter would also say that he had confidence things would work out from now on.

3. Can be expected to lecture at great length about treating others with respect and working harmoniously together.

LOW LEVEL PERFORMANCE

*Lectures or delivers ultimatums to Whipker about improving his relationships with others, or about changing his "attitude" toward people or problems.
*Tells Whipker that he is to blame for the Tech Services conflict, and demands that Whipker stop interfering in the department.
*Fails to offer his help in resolving Whipker's conflicts with Valva, subordinates, etc.

What a low level performer might do:

2. In response to Whipker's complaints about Valva, would expect this Baxter to state that Valva's department seemed to be running pretty well. He would also be expected to argue at length about how competent Valva was.

1. This Baxter can be expected to tell Whipker in no uncertain terms that he does not tolerate dissension in his ranks, and that Whipker is not to mess up the Tech Services Department.

DIMENSION F: FOSTERING PROFESSIONAL GROWTH

Offering to help Whipker identify and reach his professional goals; setting another meeting with Whipker at which they will develop an action plan for Whipker's development; recommending some preliminary actions VERSUS displaying little or no interest in Whipker's professional growth; gives poor or inappropriate advice regarding Whipker's development.

HIGH LEVEL PERFORMANCE

*Displays considerable interest in Whipker's professional development. Helps to identify problem areas, and provides appropriate developmental suggestions.
*Makes commitments to help personally in Whipker's development.
*Treats this interview as the first in a series of meetings to plot Whipker's ongoing growth.

What a high level performer might do:

7. This Baxter can be expected to suggest that he and Whipker jointly develop a list of Whipker's training needs, and then formulate a schedule of courses, seminars and independent work that can begin to address those needs. This Baxter can also be expected to schedule regular meetings at which he and Whipker can review Whipker's progress, as well as discussing any problem areas.

6. Would expect this Baxter to tell Whipker that he should try to temper his tough attitude, and would offer to attend the Dale Carnegie Course with Whipker.

AVERAGE PERFORMANCE

*Provides general developmental suggestions, and good advice on choosing courses, but fails to make a personal commitment to aid in Whipker's professional development.
*Shows moderate interest in Whipker's development; may direct Whipker to other sources within the company for developmental suggestions.

What an average performer might do:

5. This Baxter would suggest that Whipker obtain a list of courses from the personnel department, and would offer his help in choosing appropriate courses.

4. Can be expected to offer Whipker help in his general development.

3. This Baxter would direct Whipker to take a personnel management course, but not talk to him at all about what he could expect to gain from such a course, nor what the next developmental step would be.

LOW LEVEL PERFORMANCE

*Fails to offer developmental suggestions or provides poor advice regarding Whipker's professional growth.
*Expresses little or no interest in Whipker's professional development.

What a low level performer might do:

2. This Baxter could be expected to state that Whipker would have to work on his own to accomplish changes in his style.

1. If Whipker asked this Baxter for a list of things he could improve upon in order to be promoted, would expect him to be unable to provide any guidance, and to state that he didn't believe in training and development anyway.

DIMENSION G: INCREASING/MAINTAINING WHIPKER'S MOTIVATION

Supporting Whipker's excellent past performance; providing incentives for Whipker to stay at GCI and to perform effectively VERSUS failing to make commitments encouraging Whipker's continued top performance; providing little or no incentive for Whipker to remain at GCI and perform effectively; undermining Whipker's image of himself as an excellent engineer.

HIGH LEVEL PERFORMANCE

96

•Provides encouragement and appropriate incentives to persuade Whipker to remain with GCI and to strive to improve his job performance.
•Uses appropriate compliments of Whipker's technical expertise and excellent past performance to motivate Whipker to meet the objectives of the department.

What a high level performer might do:

7. This Baxter can be expected to tell Whipker that he and GCI need Whipker because of his impressive expertise and proven ability to get the job done. Can also expect him to express empathy with Whipker's job frustrations, to ask Whipker's support in terms of continued top performance,and to pledge in a sincere manner to do all he can to get Whipker more support in his present job.

6. At the end of the interview, would expect this Baxter to reiterate the commitments he has made to Whipker with regard to inquiring about job openings in higher level technical positions within GCI, to stress that Whipker is not "dead ended" in his current position, and to suggest that Whipker's excellent job performance will increase his chances of advancement.

125

AVERAGE PERFORMANCE

•Compliments Whipker appropriately but is only moderately effective in using these compliments to encourage high performance, loyalty to GCI, etc.
•Provides some incentives for Whipker to perform well and to stay at GCI, but generally makes few personal commitments to support Whipker in his job.

What an average performer might do:

5. Would expect this Baxter to compliment Whipker's past job performance, and to encourage Whipker to continue his excellent work.

4. Throughout the interview, this Baxter can be expected to emphasize his desire to keep Whipker in the company.

3. This Baxter could be expected to tell Whipker to "keep plugging" on his job because GCI needs his expertise.

LOW LEVEL PERFORMANCE

•Fails to express support for Whipker's past performance. Seems unaware of Whipker's contributions or technical expertise.
•Provides little or no incentive for Whipker to remain at GCI.

What a low level performer might do:

2. Can be expected to tell Whipker he appears to doing an adequate job in his department but that he could probably be doing better.

1. After discussing Whipker's problems within GCI, this Baxter would suggest that Whipker leave the company since he was so dissatisfied.

126

APPENDIX J: POST-EXPERIMENTAL QUESTIONNAIRE

Biographical Information

Please complete the following short questionnaire. The information will be used in conjunction with the experimental materials you complete for this study. Obviously, this information is confidential, and we will treat it as such. Please return this form to the front of the room after you have completed it. Thank you.

Name_____(please print)

Social Security # _____

1. Sex _____
2. Age _____
3. Total years of full-time work experience (if any) _____
4. Total years experience as supervisor/manager (if any) _____
5. If you have completed performance appraisals for employees under your supervision, what is the approximate number you have done to date? _____
6. If you have provided feedback interviews on employees' performance, what is the approximate number to date?_____
7. Have you ever been a supervisor for engineers? If yes, for how many years? _____

Opinions on Performance Appraisal

Before beginning the items on this questionnaire, please answer the following question by circling the correct response:

What was the purpose of the performance ratings, that is, what are they to be used for?

a. part of a promotion decision
b. for research in validating a selection battery
c. for personal growth and development of the individuals
d. I don't know

In the remainder of this questionnaire, there are various items that ask your opinion about performance appraisal. The questions are concerned with your opinions about the performance ratings you have just made in this study. Answer each of the items in this questionnaire using the following scale. Place the number which corresponds to your answer in the blank space beside the question.

1 Not at all
2 To a little extent
3 To a moderate extent
4 To a great extent
5 To a very great extent

_____ 1. To what extent do you believe that the true purpose of this study was the one explained by the researcher?

_____ 2. To what extent are you confident we could use your ratings to evaluate test scores (validate) on the employees depicted in the job situations?

_____ 3. To what extent do you feel you could defend your ratings to the others in your group who gave different ratings to the same employees in the job situations?

_____ 4. To what extent did the directions for using the rating scale help you to use it properly?

_____ 5. To what extent do you think other persons in this study gave higher ratings to help out the persons depicted in the job situations?

1 Not at all
2 To a little extent
3 To a moderate extent
4 To a great extent
5 To a very great extent

_____ 6. To what extent did you "inflate" your ratings to give the employees in the job situations a higher score?

_____ 7. To what extent do you trust that the performance ratings you made are going to be used for the specific purpose described in the study?

_____ 8. To what extent do you feel your ratings accurately captured the true performance of the people you rated?

_____ 9. Overall, to what extent did you feel confident about the ratings you made in this study?

_____ 10. To what extent would your closest friend describe you as a person who is overly concerned with accuracy in your work performance?

_____ 11. To what extent were you clear as to the standards to use in making your ratings?

_____ 12. To what extent did you understand what the ratings as described on the ratings form were trying to measure?

_____ 13. To what extent were you uncomfortable giving negative ratings to the employees in the job situations?

_____ 14. To what extent would you be willing to participate in another study of this kind in the future?

_____ 15. To what extent did you understand how to complete the performance ratings?

_____ 16. To what extent do you feel the results of this study will provide information to the sponsoring organization?

_____ 17. To what extent did you really make an "extra effort" to carefully pay attention to the job performance materials in order to make your ratings accurate?

1 Not at all
2 To a little extent
3 To a moderate extent
4 To a great extent
5 To a very great extent

_____ 18. To what extent would you describe yourself as being overly concerned with details in other aspects of your life?

_____ 19. To what extent did you enjoy completing the performance ratings in this study?

_____ 20. To what extent do you feel the ratings you completed in this study are going to be useful and worthwhile for this research?

_____ 21. To what extent do you feel you were able to accurately distinguish between good and poor performers in the job situations?

_____ 22. To what extent do you feel the results of this study will be useful for application in real organizations?

_____ 23. To what extent are you confident that we could use your ratings as the performance measures for a $100,000 selection project?

_____ 24. To what extent do you feel this was a useful study?

_____ 25. Given the circumstances of this study, to what extent were you very willing to complete the ratings?

_____ 26. To what extent do you feel the rating form used in this study is a good one?

_____ 27. To what extent did you feel it was important for you to make accurate ratings in this study?

_____ 28. Based on your experience in this study, how important is it to you to make any performance ratings you do in the future as accurate as you can?

1 Not at all
2 To a little extent
3 To a moderate extent
4 To a great extent
5 To a very great extent

_____ 29. To what extent did the rating form used in this study enable you to make accurate ratings?

_____ 30. To what extent are you satisfied you made the most accurate ratings you could in this study?

_____ 31. To what extent did the rating form enable you to evaluate the performance of the employees in the job situations fairly?

_____ 32. To what extent do you feel the written instructions with the rating form were completely clear?

Thank you very much for your help in completing this study. If you are willing to participate in another performance rating study similar to this one in the future, please print your name, local address, and phone number below. We will be in touch with you sometime during the Spring semester.

Name: _____

Address: _____

_____

Phone: _____